

# Regularization, sparse recovery, and median-of-means tournaments \*

Gábor Lugosi<sup>†‡§</sup>

Shahar Mendelson<sup>¶</sup>

January 17, 2017

## Abstract

A regularized risk minimization procedure for regression function estimation is introduced that achieves near optimal accuracy and confidence under general conditions, including heavy-tailed predictor and response variables. The procedure is based on median-of-means tournaments, introduced by the authors in [8]. It is shown that the new procedure outperforms standard regularized empirical risk minimization procedures such as LASSO or SLOPE in heavy-tailed problems.

**2010 Mathematics Subject Classification:** 62J02, 62G08, 60G25.

## 1 Introduction

### 1.1 Empirical risk minimization, regularization

Regression function estimation is a fundamental problem in statistics and machine learning. In the most standard formulation of the problem,  $(X, Y)$  is a pair of random variables in which  $X$ , taking values in some general measurable space  $\mathcal{X}$ , represents the observation (or feature vector) and one would like to approximate the unknown real value  $Y$  by a function of  $X$ . In other words, one is interested in finding a function  $f : \mathcal{X} \rightarrow \mathbb{R}$  such that  $f(X)$  is “close” to  $Y$ . As the vast majority of the literature, we measure the quality of  $f$  by the *risk*

$$R(f) = \mathbb{E}(f(X) - Y)^2,$$

which is well defined whenever  $f(X)$  and  $Y$  are square integrable, assumed throughout the paper. Clearly, the best possible function is the *regression function*  $m(X) = \mathbb{E}(Y|X)$ .

However, in statistical problems, the joint distribution of  $(X, Y)$  is unknown and the regression function is impossible to compute. Instead, a sample  $\mathcal{D}_N = ((X_1, Y_1), \dots, (X_N, Y_N))$  of independent copies of the pair  $(X, Y)$  is available.

---

\*Gábor Lugosi was supported by the Spanish Ministry of Economy and Competitiveness, Grant MTM2015-67304-P and FEDER, EU. Shahar Mendelson was supported in part by the Israel Science Foundation.

<sup>†</sup>Department of Economics and Business, Pompeu Fabra University, Barcelona, Spain, gabor.lugosi@upf.edu

<sup>‡</sup>ICREA, Pg. Llus Companys 23, 08010 Barcelona, Spain

<sup>§</sup>Barcelona Graduate School of Economics

<sup>¶</sup>Department of Mathematics, Technion, I.I.T, and Mathematical Sciences Institute, The Australian National University, shahar@tx.technion.ac.il

A popular and thoroughly studied approach is to select a function  $\hat{f}_N$  from a fixed class  $\mathcal{F}$  of functions. Formally, a *learning procedure* is a map  $\Phi : (\mathcal{X} \times \mathbb{R})^N \rightarrow \mathcal{F}$  that assigns to each sample  $\mathcal{D}_N = (X_i, Y_i)_{i=1}^N$  a (random) function  $\Phi(\mathcal{D}_N) = \hat{f}_N$ .

If the class  $\mathcal{F}$  is sufficiently “large,” then it is reasonable to expect that the best function in the class

$$f^* = \operatorname{argmin}_{f \in \mathcal{F}} \mathbb{E}(f(X) - Y)^2$$

has an acceptable performance. We assume throughout that the minimum is attained and  $f^* \in \mathcal{F}$  is unique. In fact, we assume that  $\mathcal{F}$  is a closed and convex subset of  $L_2(\mu)$ —where  $\mu$  denotes the distribution of  $X$ —, guaranteeing the existence and uniqueness of  $f^*$ .

The quality of a learning procedure is typically measured by the *mean squared error*

$$\|\hat{f}_N - f^*\|_{L_2}^2 = \mathbb{E}((\hat{f}_N(X) - f^*(X))^2 | \mathcal{D}_N) ,$$

where, for  $q \geq 1$ , we use the notation

$$\|f - g\|_{L_q} = (\mathbb{E}|f(X) - g(X)|^q)^{1/q} \quad \text{and also} \quad \|f - Y\|_{L_q} = (\mathbb{E}|f(X) - Y|^q)^{1/q} .$$

A closely related, though not equivalent, measure of performance is the *excess risk*, defined by the conditional expectation

$$R(\hat{f}_N) - R(f^*) = \mathbb{E}((\hat{f}_N(X) - Y)^2 | \mathcal{D}_N) - \mathbb{E}(f^*(X) - Y)^2 .$$

The goal of the statistical learning problem is to find a learning procedure that achieves a good *accuracy* with a high *confidence*. In particular, for  $r > 0$  and  $\delta \in (0, 1)$ , we say that a procedure achieves accuracy  $r$  with confidence  $1 - \delta$  (e.g., for the mean squared error) if

$$\mathbb{P}\left(\|\hat{f}_N - f^*\|_{L_2} \leq r\right) \geq 1 - \delta .$$

High accuracy and high confidence (i.e., small  $r$  and small  $\delta$ ) are obviously conflicting requirements. The achievable tradeoff has been thoroughly studied and it is fairly well understood. We refer the reader to Lecué and Mendelson [4], Lugosi and Mendelson [8] for recent accounts.

The most standard approach for a learning procedure is *empirical risk minimization* (ERM), also known as *least squares regression* in which

$$\hat{f}_N \in \operatorname{argmin}_{f \in \mathcal{F}} \sum_{i=1}^N (f(X_i) - Y_i)^2$$

(where we assume that the minimum is achieved). It is now well understood (see, e.g., Lecué and Mendelson [4]) that if the distribution of  $Y$ ,  $f(X)$  and  $(f - h)(X)$  for  $f, h \in \mathcal{F}$  all have well-behaved tails (sub-Gaussian in a certain, strong, sense), then empirical risk minimization achieves nearly optimal accuracy and the best possible confidence for that level of accuracy. On the other hand, when either  $Y$  or  $(f - h)(X)$  for  $f, h \in \mathcal{F} \cup \{0\}$  may have heavier tails, empirical risk minimization suffers, as atypical values in the sample distort the empirical means. Indeed, in such cases significantly better learning procedures exist, as it was recently pointed out by Lugosi and Mendelson [8].

Even for well-behaved distributions, the accuracy achievable by empirical risk minimization is only acceptable if the class  $\mathcal{F}$  is small, otherwise *overfitting* becomes inevitable. A common way of avoiding overfitting is by *regularization*.

In regularized risk minimization one gives priority to functions in  $\mathcal{F}$  according to some prior belief of “simplicity”. More precisely, let  $\Psi$  be a norm defined on a vector space  $E$  containing  $\mathcal{F}$ . A small value of  $\Psi(f)$  is interpreted as simplicity and simple functions are given priority by way of adding a penalty term to the empirical risk that is proportional to  $\Psi(f)$ . In particular, for some *regularization parameter*  $\lambda > 0$ , a regularized risk minimizer selects

$$\hat{f}_N \in \operatorname{argmin}_{f \in \mathcal{F}} \left( \frac{1}{N} \sum_{i=1}^N (f(X_i) - Y_i)^2 + \lambda \Psi(f) \right).$$

The term  $\Psi(f)$  is sometimes called the *penalty*. In some applications, notably in ridge regression, the penalty is not a norm but rather a squared norm. However, in this paper we focus on norm penalties that encompass various notions of sparsity, for example, the LASSO and SLOPE, discussed in detail below.

## 1.2 Why run tournaments?

The problem that motivated this work is the better understanding of the tradeoff between accuracy and confidence in regularized learning procedures. Since regularized procedures require the minimization of a functional that has the empirical mean as a component, they suffer from the same disadvantages as empirical risk minimization. These are highly visible when either class members  $f(X)$  or the target  $Y$  are heavy tailed in some sense. Thus, the application of regularized ERM in heavy-tailed problems results in a suboptimal tradeoff between the accuracy with which the procedure performs and the confidence with which that accuracy can be guaranteed. A typical sample contains a large subset of atypical points and misleads the empirical minimization procedure.

The idea of a median-of-means tournament was introduced in [8] to address the issue of samples that may contain many atypical points. It turns out that this learning procedure attains the optimal accuracy/confidence tradeoff under rather minimal conditions and, in particular, in heavy-tailed problems.

In what follows we study the optimal tradeoff for problems usually addressed using regularization procedures.

### Example: sparse recovery

It is instructive to keep in mind the important example of sparse-recovery in  $\mathbb{R}^d$ . Consider the following standard setup: Let  $X$  be an isotropic random vector in  $\mathbb{R}^d$  (that is, for every  $t \in \mathbb{R}^d$ ,  $\mathbb{E} \langle t, X \rangle^2 = \|t\|_2^2$ , where  $\|\cdot\|_2$  denotes the Euclidean norm in  $\mathbb{R}^n$ ). Let  $Y$  be the unknown target random variable and set  $t^*$  to be the minimizer in  $\mathbb{R}^d$  of the risk functional  $t \rightarrow \mathbb{E} (\langle X, t \rangle - Y)^2$ . Thus,  $\mathcal{F}$  contains all linear functions  $f(x) = \langle x, t \rangle$  for  $t \in \mathbb{R}^d$ . In sparse recovery problems one believes that  $t^*$  is supported on at most  $s$  coordinates with respect to the standard basis in  $\mathbb{R}^d$ —or at least it is well approximated by an  $s$ -sparse vector—but one does not know that for certain. The LASSO procedure (introduced in [15]) selects  $\hat{t} \in \mathbb{R}^d$  that minimizes the regularized empirical squared-loss functional

$$t \rightarrow \frac{1}{N} \sum_{i=1}^N (\langle t, X_i \rangle - Y_i)^2 + \lambda \|t\|_1$$

for a well chosen regularization parameter  $\lambda$ , and  $\|t\|_1 = \sum_{i=1}^d |t_i|$  is the  $\ell_1$ -norm of  $t$ .

It turns out (see Lecué and Mendelson [5]) that if  $X$  is sufficiently sub-Gaussian then one may obtain nontrivial estimates on the performance of the LASSO. The quantities  $c_i(\cdot)$  denote positive constants that depend on the argument only.

**Theorem 1.1.** (*Lecué and Mendelson [5].*) *Let  $(X, Y)$  be as described above and set  $0 < \delta < 1$ . Assume that there is some  $v \in \mathbb{R}^d$  supported on at most  $s$  coordinates for which*

$$\|t^* - v\|_1 \leq c_1(\delta) \|\xi\|_{L_q} s \sqrt{\frac{\log(ed)}{N}}.$$

*If  $\lambda = c_2(L, \delta) \|\xi\|_{L_q} \sqrt{\log(ed)/N}$  and  $N \geq c_3(L) s \log(ed/s)$  then, with probability at least  $1 - \delta$ , the LASSO estimator with regularization parameter  $\lambda$  satisfies that*

$$\|\hat{t} - t^*\|_2 \leq c_4(L, \delta) \|\xi\|_{L_q} \sqrt{s} \cdot \sqrt{\frac{\log(ed)}{N}}$$

*and*

$$\|\hat{t} - t^*\|_1 \leq c_4(L, \delta) \|\xi\|_{L_q} s \sqrt{\frac{\log(ed)}{N}}.$$

Note that the results of Theorem 1.1 hold even when  $t^*$  is not  $s$ -sparse, but rather, when  $t^*$  is only approximated by an  $s$ -sparse vector.

Thus, the LASSO can identify the degree of sparsity of  $t^*$  and perform (almost) as if it had been given the information of the degree of sparsity of  $t^*$ . However, if  $\langle t^*, X \rangle - Y$  happens to be heavy-tailed then the confidence with which the above accuracy can be attained is rather weak:  $c(\delta)$  and  $c(\delta, L)$  depend polynomially on  $1/\delta$ .

The purpose of this paper is to introduce a learning procedure — a *regularized median of means tournament* — that performs as well as regularized empirical risk minimization but under much more general conditions on the distribution. In fact, its performance corresponds to that of regularized empirical risk minimization in sub-Gaussian problems, even though the problem at hand may be heavy tailed. For example, it is not surprising that the accuracy/confidence tradeoff of the LASSO deteriorates when the problem is heavy-tailed, because the LASSO is based on minimization of the empirical risk. In contrast, the regularized procedure we introduce performs well even if we drop the sub-Gaussian assumptions and replace them by considerably weaker ones.

Our benchmark is a general estimate due to Lecué and Mendelson [5] on the performance of regularized empirical risk minimization, and the main result of this paper (Theorem 4.3 below) parallels the main finding of [5]. The proof of Theorem 4.3 combines the techniques of [5] with those of [8]. We then use the LASSO as a proof of concept and show that the procedure we propose matches the performance of the LASSO in the well-behaved case even when the problem is heavy tailed. We also work out a procedure analogous to SLOPE and present similar findings.

### 1.3 Two examples

Before we turn to a presentation of the regularized tournament procedure and the technical machinery we require, let us describe in more detail the two applications mentioned previously, both of which originating in sparse recovery: LASSO, and SLOPE (see, [1, 3, 14, 15]). The two are regularized empirical risk minimization procedures in  $\mathbb{R}^d$ , and we would like to compare their performance with that of the regularized tournament procedure we introduce below.

As we mentioned, the LASSO is a regularized empirical risk minimization procedure that uses the regularization function  $\Psi(t) = \|t\|_1$ , while the regularization function of SLOPE is defined using a set of non-increasing weights  $(\beta_i)_{i=1}^d$ . The corresponding norm is

$$\Psi(t) = \sum_{i=1}^d \beta_i t_i^*,$$

where  $(t_i^*)_{i=1}^d$  denotes the non-increasing rearrangement of  $(|t_i|)_{i=1}^d$ . Clearly, SLOPE is a generalized version of LASSO, as the latter is given by the choice  $\beta_i = 1$  for  $1 \leq i \leq d$ . LASSO and SLOPE exhibit an almost miraculous feature: although the regularization functions have little to do with sparsity, regularized empirical risk minimization preforms extremely well when the true minimizer  $t^*$  is sparse, or if it is at least well-approximated by a sparse vector.

Many results are known on the performance of LASSO and SLOPE, and almost all of them hold only when both the random vector  $X$  and the target  $Y$  have well behaved tails. One exception is [5] that studies the case of a potentially heavy-tailed  $Y$ , with the resulting estimate outlined in Theorem 1.1. It shows that while a high degree of accuracy is possible—the same as if the problem were purely Gaussian, the confidence is rather weak. That weak confidence is caused by the heavy tail of  $\xi$ , the insufficient sub-Gaussian moments of linear forms, and the fact that empirical minimization procedures are highly sensitive to atypical sample points. It turns out that a similar performance bound holds for SLOPE with the choice of weights  $\beta_i \leq C\sqrt{\log(ed/i)}$ .

**Theorem 1.2.** (*Lecué and Mendelson [5].*) *There exist constants  $c_1, c_2$  and  $c_3$  that depend only on  $L, \delta$  and  $C$  for which the following holds.*

*Let  $\Psi(t)$  denote the SLOPE norm for the weights  $(\beta_i)_{i=1}^d$ . If there is  $v \in \mathbb{R}^d$  that satisfies  $|\text{supp}(v)| \leq s$  and*

$$\Psi(t^* - v) \leq c_1 \|\xi\|_{L_q} \frac{s}{\sqrt{N}} \log \left( \frac{ed}{s} \right),$$

*then for  $N \geq c_2 s \log(ed/s)$  and with the choice of  $\lambda = c_2 \|\xi\|_{L_q} / \sqrt{N}$ , one has*

$$\Psi(\hat{t} - t^*) \leq c_3 \|\xi\|_{L_q} \frac{s}{\sqrt{N}} \log \left( \frac{ed}{s} \right) \quad \text{and} \quad \|\hat{t} - t^*\|_2 \leq c_3 \|\xi\|_{L_q} \sqrt{\frac{s}{N} \log \left( \frac{ed}{s} \right)}$$

*with probability at least  $1 - \delta$ .*

We show that median-of-means versions tournament versions of LASSO and SLOPE perform better than their regularized empirical risk minimization counterparts, simply because median-of-means is far more robust to atypical sample points than empirical minimization. This leads to the ‘Gaussian’ accuracy, but with the optimal confidence, under the same assumptions.

## The tournament LASSO

The tournament LASSO is just a regularized tournament procedure (defined accurately below) for  $\Psi(t) = \|t\|_1$ . As the next result shows, it significantly outperforms the LASSO in heavy-tailed problems.

**Theorem 1.3.** Assume that  $X$  is an isotropic random vector and that for every  $t \in \mathbb{R}^d$  and any  $1 \leq p \leq c \log d$ ,  $\|\langle t, X \rangle\|_{L_p} \leq L\sqrt{p}\|\langle t, X \rangle\|_{L_2}$ . Let  $t^* = \operatorname{argmin}_{t \in \mathbb{R}^d} \mathbb{E}(Y - \langle t, X \rangle)^2$  and assume that  $\|Y - \langle t^*, X \rangle\|_{L_4} \leq \sigma$ .

If there is  $v$  that is  $s$ -sparse such that

$$\|t^* - v\|_1 \leq c_1(L)\sigma \cdot s \sqrt{\frac{\log(ed/s)}{N}},$$

$N \geq c_2(L)s \log(ed/s)$ , and

$$\hat{r} \geq c_3(L)\sigma \sqrt{\frac{s}{N} \log\left(\frac{ed}{s}\right)},$$

then, with probability at least

$$1 - 2 \exp\left(-c_4(L)N \min\left\{1, \left(\frac{\hat{r}}{\sigma}\right)^2\right\}\right),$$

the tournament LASSO produces  $\hat{t}$  that satisfies

$$\|\hat{t} - t^*\|_2 \leq c_5(L)\hat{r} \quad \text{and} \quad \|\hat{t} - t^*\|_1 \leq c_5(L)\sigma s \sqrt{\frac{\log(ed/s)}{N}}.$$

It follows that the tournament LASSO yields the same accuracy as the standard LASSO, but with a much better confidence. Moreover, the confidence improves when we require a weaker accuracy. The crucial point is that the tournament LASSO attains the optimal accuracy/confidence tradeoff even though the problem is very far from being sub-Gaussian. Linear functionals  $\langle X, t \rangle$  exhibit a sub-Gaussian moment growth only up to  $p \sim \log d$  rather than for any  $p \geq 1$ , and the noise  $\xi = \langle t^*, X \rangle - Y$  is only assumed to belong to  $L_4$ . There is no hope that regularized empirical risk minimization would come close to this accuracy/confidence tradeoff under such assumptions, but the tournament LASSO does just that.

### The tournament SLOPE

The tournament SLOPE is simply a regularized tournament procedure in  $\mathbb{R}^d$  for the norm  $\Psi(t) = \sum_{i=1}^d \beta_i t_i^*$ , where  $\beta_i \leq C\sqrt{\log(ed/i)}$ . We obtain the following performance bound, proved in Section 6.

**Theorem 1.4.** Assume that  $X$  is an isotropic random vector and that for every  $t \in \mathbb{R}^d$  and any  $1 \leq p \leq c \log d$ ,  $\|\langle t, X \rangle\|_{L_p} \leq L\sqrt{p}\|\langle t, X \rangle\|_{L_2}$ . Let  $t^* = \operatorname{argmin}_{t \in \mathbb{R}^d} \mathbb{E}(Y - \langle t, X \rangle)^2$  and assume that  $\|Y - \langle t^*, X \rangle\|_{L_4} \leq \sigma$ . If there is  $v$  that is  $s$ -sparse such that

$$\|t^* - v\|_1 \leq c_1(L)\sigma \cdot \frac{s \log(ed/s)}{\sqrt{N}},$$

$N \geq c_2(L)s \log(ed/s)$ , and

$$\hat{r} \geq c_3(L)\sigma \sqrt{\frac{s}{N} \log\left(\frac{ed}{s}\right)},$$

then, with probability at least

$$1 - 2 \exp\left(-c_4(L)N \min\left\{1, \left(\frac{\hat{r}}{\sigma}\right)^2\right\}\right),$$

the tournament SLOPE produces  $\hat{t}$  that satisfies

$$\|\hat{t} - t^*\|_2 \leq c_5(L)\sigma \sqrt{\frac{s}{N} \log\left(\frac{ed}{s}\right)}$$

and

$$\Psi(\hat{t} - t^*) = \sum_{i=1}^d (\hat{t} - t^*)_i^* \sqrt{\log(ed/i)} \leq c_5(L)\sigma \frac{s}{\sqrt{N}} \log\left(\frac{ed}{s}\right).$$

Therefore, one can obtain the same accuracy as the standard SLOPE, but with a much better confidence. The confidence improves for a weaker accuracy, and just like the tournament LASSO, the tournament SLOPE displays the optimal accuracy/confidence tradeoff, but allowing heavy-tailed distributions.

## 2 Assumptions, background

In the general setup we study, we merely assume a rather weak fourth-moment assumption. More precisely, we work under the following conditions.

**Assumption 2.1.** *Let  $\mathcal{F} \subset L_2(\mu)$  be a locally compact, convex class of functions. Let  $Y \in L_2$  and assume that, for some constant  $L > 0$ ,*

- *for every  $f, h \in \mathcal{F}$ ,  $\|f - h\|_{L_4} \leq L\|f - h\|_{L_2}$ ;*
- *$\|f^* - Y\|_{L_4} \leq \sigma_4$  for a known value  $\sigma_4$ .*

**Remark 2.1.** *The second condition may easily be replaced by a combination of two assumptions: that for every  $f \in \mathcal{F}$ ,  $\|f - Y\|_{L_4} \leq L\|f - Y\|_{L_2}$ ; and that  $\|f^* - Y\|_{L_2} \leq \sigma$  for some known constant  $\sigma > 0$ . Also, in the case of independent additive noise, that is, when  $Y = f_0(X) + W$ , the assumption that  $\|f^* - Y\|_{L_4} \leq \sigma_4$  may be replaced by the minimal one, that  $\|W\|_{L_2} \leq \sigma$ . The necessary modifications to the proofs are straightforward and we do not explore this observation further.*

### 2.1 Complexity parameters of a class

Before describing the regularized median of mean tournament, we introduce four parameters of “complexity” depending both on the class  $\mathcal{F}$  and the distribution of  $(X, Y)$ . These complexity parameters are essential in describing the optimal performance of learning procedures. For detailed discussion on the meaning and role of these parameters, we refer to Mendelson [9, 10] and Lugosi and Mendelson [8]. As explained in those papers, the relevant complexity of a class  $\mathcal{F}$  has four components, reflected in four parameters  $\lambda_{\mathbb{Q}}, \lambda_{\mathbb{M}}, r_E, \tilde{r}_{\mathbb{M}}$  that we define next.

First we need some notation. Denote the unit ball in  $L_2(\mu)$  by  $D = \{f : \|f\|_{L_2} \leq 1\}$  and let  $S = \{f : \|f\|_{L_2} = 1\}$  be the unit sphere. For  $h \in L_2(\mu)$  and  $r > 0$ , we write  $D_h(r) = \{f : \|f - h\|_{L_2} \leq r\}$ . In a similar fashion for the norm  $\Psi$  used as a regularization function, let  $\mathcal{B} = \{f : \Psi(f) \leq 1\}$ , set  $\rho\mathcal{B} = \{f : \Psi(f) \leq \rho\}$  and  $\mathcal{B}_h(\rho) = \{f : \Psi(f - h) \leq \rho\}$ .

Define the star-shaped hull of the class  $\mathcal{F}$  centred in  $h$  by

$$\text{star}(\mathcal{F}, h) = \{\lambda f + (1 - \lambda)h : 0 \leq \lambda \leq 1, f \in \mathcal{F}\}.$$



In what follows we make two important modifications to the definitions of the complexity parameters used in [9, 10, 8]. First, just like in the above-mentioned articles, we are interested in “localized” classes. However, because regularized procedures are affected by two norms,  $\Psi$  and  $L_2(\mu)$ , the localization has to be with respect to both of them. Therefore, the “localization” of  $\mathcal{F}$ , centred in  $h$  and of radii  $\rho, r > 0$  is defined by

$$\mathcal{F}_{h,\rho,r} = \text{star}(\mathcal{F} - h, 0) \cap (\rho\mathcal{B} \cap rD) .$$

Clearly, if  $\mathcal{F}$  is convex then for any  $h \in \mathcal{F}$ ,  $\text{star}(\mathcal{F} - h, 0) = \mathcal{F} - h$  and

$$\mathcal{F}_{h,\rho,r} = \{f - h : f \in \mathcal{F}, \Psi(f - h) \leq \rho, \|f - h\|_{L_2} \leq r\} = (\mathcal{F} - h) \cap (\rho\mathcal{B} \cap rD) .$$

The second minor modification is that each complexity parameter is associated with the ‘worse case’ centre  $h \in \mathcal{F}'$  for some fixed  $\mathcal{F}' \subset \mathcal{F}$ , and not necessarily with the whole of  $\mathcal{F}$ .

Two of the four parameters are defined using the notion of *packing numbers*.

**Definition 2.2.** *Given a set  $H \subset L_2(\mu)$  and  $\varepsilon > 0$ , denote the  $\varepsilon$ -packing number of  $H$  by  $\mathcal{M}(H, \varepsilon D)$ . In other words,  $\mathcal{M}(H, \varepsilon D)$  is the maximal cardinality of a subset  $\{h_1, \dots, h_m\} \subset H$ , for which  $\|h_i - h_j\|_{L_2} \geq \varepsilon$  for every  $i \neq j$ .*

The first relevant parameter  $\lambda_{\mathbb{Q}}$  is defined as follows with appropriate numerical constants  $\kappa$  and  $\eta$ :

**Definition 2.3.** *Fix  $\rho > 0$  and  $h \in \mathcal{F}$ . For  $\kappa, \eta > 0$ , set*

$$\lambda_{\mathbb{Q}}(\kappa, \eta, h, \rho) = \inf\{r : \log \mathcal{M}(\mathcal{F}_{h,\rho,r}, \eta r D) \leq \kappa^2 N\} . \quad (2.1)$$

For  $\mathcal{F}' \subset \mathcal{F}$  let

$$\lambda_{\mathbb{Q}}(\kappa, \eta, \rho) = \sup_{h \in \mathcal{F}'} \lambda_{\mathbb{Q}}(\kappa, \eta, h, \rho) .$$

While  $\kappa$  and  $\eta$  are adjustable parameters, we are mainly interested in the behaviour of  $\lambda_{\mathbb{Q}}$  as a function of  $\rho$ . The way one selects  $\rho$  is clarified later.

The next parameter, denoted by  $\lambda_{\mathbb{M}}$ , is also defined in terms of the packing numbers of the localization  $\mathcal{F}_{h,\rho,r}$ , though at a different scaling than  $\lambda_{\mathbb{Q}}$ .

**Definition 2.4.** *Fix  $h \in \mathcal{F}$  and  $\rho > 0$ . Let  $\kappa > 0$ ,  $0 < \eta < 1$ , and define*

$$\lambda_{\mathbb{M}}(\kappa, \eta, h, \rho) = \inf\{r : \log \mathcal{M}(\mathcal{F}_{h,\rho,r}, \eta r D) \leq \kappa^2 N r^2\} . \quad (2.2)$$

Also, for  $\mathcal{F}' \subset \mathcal{F}$  let

$$\lambda_{\mathbb{M}}(\kappa, \eta, \rho) = \sup_{h \in \mathcal{F}'} \lambda_{\mathbb{M}}(\kappa, \eta, h, \rho) .$$

For the remaining two complexity parameters, let  $(\varepsilon_i)_{i=1}^N$  be independent, symmetric  $\{-1, 1\}$ -valued random variables that are independent of  $(X_i, Y_i)_{i=1}^N$ .

**Definition 2.5.** *Fix  $h \in \mathcal{F}$  and  $\rho > 0$ . For  $\kappa > 0$  let*

$$r_E(\kappa, h, \rho) = \inf \left\{ r : \mathbb{E} \sup_{u \in \mathcal{F}_{h,\rho,r}} \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N \varepsilon_i u(X_i) \right| \leq \kappa \sqrt{N} r \right\} , \quad (2.3)$$

and for  $\mathcal{F}' \subset \mathcal{F}$  set  $r_E(\kappa, \rho) = \sup_{h \in \mathcal{F}'} r_E(\kappa, h, \rho)$ .



**Definition 2.6.** Fix  $h \in \mathcal{F}$  and  $\rho > 0$ . For  $\kappa > 0$ , set  $\bar{r}_{\mathbb{M}}(\kappa, h, \rho)$  to be

$$\bar{r}_{\mathbb{M}}(\kappa, h, \rho) = \inf \left\{ r : \mathbb{E} \sup_{u \in \mathcal{F}_{h, \rho, r}} \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N \varepsilon_i u(X_i) \cdot (h(X_i) - Y_i) \right| \leq \kappa \sqrt{N} r^2 \right\}. \quad (2.4)$$

For  $\sigma > 0$  put  $\mathcal{F}_Y^{(\sigma)} = \{f \in \mathcal{F}' : \|f(X) - Y\|_{L_2} \leq \sigma\}$  and let  $\tilde{r}_{\mathbb{M}}(\kappa, \sigma, \rho) = \sup_{h \in \mathcal{F}_Y^{(\sigma)}} \bar{r}_{\mathbb{M}}(\kappa, h, \rho)$ .

Finally, suppose that the distribution of  $(X, Y)$  is such that  $\|Y - f^*(X)\|_{L_4} \leq \sigma_4$  for a known constant  $\sigma > 0$ . The ‘‘complexity’’ of  $\mathcal{F}$  relative to centres in  $\mathcal{F}'$  and radius  $\rho$  is

$$r^*(\mathcal{F}, \mathcal{F}', \rho) = \max\{\lambda_{\mathbb{Q}}(c_1, c_2, \rho), \lambda_{\mathbb{M}}(c_1/\sigma_4, c_2, \rho), r_E(c_1, \rho), \tilde{r}_{\mathbb{M}}(c_1, \sigma_4, \rho)\}. \quad (2.5)$$

Here  $c_1, c_2$  are appropriate positive numerical constants. (‘‘Appropriate’’ means that  $r^*(\mathcal{F}, \mathcal{F}', \rho)$  satisfies Propositions 5.1, 5.5 and 5.7 below). The existence of such constants is proved in [8] when  $\mathcal{F}' = \mathcal{F}$ , i.e., when any function in  $\mathcal{F}$  is a ‘legal choice’ of a centre, and under Assumption 2.1. In that case, the constants depend only on the value of  $L$ .

When  $\mathcal{F}$  and  $\mathcal{F}'$  are clear from the context, we simply write  $r^*(\rho)$  for  $r^*(\mathcal{F}, \mathcal{F}', \rho)$ .

### Example: linear regression with $\ell_1$ regularization

To give the reader some feeling of the nature of  $r^*(\rho)$  and the type of estimates that are needed for its identification, let  $\mathcal{F}$  be a vector space. Hence, for any  $h \in \mathcal{F}$ ,  $\mathcal{F} - h = \mathcal{F}$  and thus

$$\mathcal{F}_{h, \rho, r} = \mathcal{F} \cap \rho \mathcal{B} \cap rD.$$

In particular,  $\mathcal{F}_{h, \rho, r}$  is independent of the choice of centre  $h$  and there is no ‘diversity’ in the localized sets one encounters. Also, if  $\mathcal{F}$  is convex and centrally symmetric (i.e., if  $f \in \mathcal{F}$  then  $-f \in \mathcal{F}$ ), the richest localized set is essentially when the centre is  $h = 0$ . Indeed, in such a case,

$$\mathcal{F}_{h, \rho, r} = (\mathcal{F} - h) \cap (\rho \mathcal{B} \cap rD) \subset 2\mathcal{F} \cap (\rho \mathcal{B} \cap rD),$$

which is a localization of  $2\mathcal{F}$  for  $h = 0$ . Thus, when  $\mathcal{F}$  is centrally symmetric, it suffices to study its localizations associated with a single centre  $-h = 0$ .

In the case of the LASSO  $\mathcal{F} = \{\langle t, \cdot \rangle : t \in \mathbb{R}^d\}$  and therefore  $\mathcal{F}$  is a vector space consisting of linear functionals. From here on we identify the linear functional  $\langle t, \cdot \rangle$  with  $t \in \mathbb{R}^d$ , and thus identify  $\mathcal{F}$  with  $\mathbb{R}^d$ . The regularization function used in the LASSO is  $\Psi(t) = \|t\|_1$  and  $t^* = \operatorname{argmin}_{t \in \mathbb{R}^d} \mathbb{E}(\langle t, X \rangle - Y)^2$ .

Using our notation,  $D = \{t \in \mathbb{R}^d : \mathbb{E} \langle X, t \rangle^2 \leq 1\}$ , and if  $X$  is isotropic, then  $D = B_2^d$ , the Euclidean unit ball in  $\mathbb{R}^d$  (though, in general,  $D$  is an ellipsoid in  $\mathbb{R}^d$ ). Hence, all the localizations associated with the LASSO are

$$\mathcal{F}_{t, \rho, r} = \rho B_1^d \cap rD,$$

where  $B_1^d = \{t \in \mathbb{R}^d : \|t\|_1 \leq 1\}$ . Moreover, if  $X$  is isotropic then

$$\mathcal{F}_{t, \rho, r} = \rho B_1^d \cap rB_2^d.$$

It follows that if one is to identify  $\lambda_{\mathbb{Q}}$  and  $\lambda_{\mathbb{M}}$ , it suffices to obtain estimates on the Euclidean packing numbers

$$\log \mathcal{M}(\rho B_1^d \cap rB_2^d, \eta r B_2^d)$$

and, in order to bound  $r_E$  and  $\bar{r}_M$  it suffices to control the expectations of the empirical processes

$$\mathbb{E} \sup_{u \in \rho B_1^d \cap r B_2^d} \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N \varepsilon_i \langle u, X_i \rangle \right|$$

and

$$\mathbb{E} \sup_{u \in \rho B_1^d \cap r B_2^d} \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N \varepsilon_i \xi_i \langle u, X_i \rangle \right|$$

respectively, where in the latter, the multipliers are  $\xi_i = \langle t^*, X_i \rangle - Y_i$ .

The question of obtaining sharp bounds on these complexity parameters, even for the LASSO, is nontrivial, and a rich mathematical theory has been developed just for that goal (see, e.g., the books [7, 16, 2], and for more recent results, [13]). We study the parameters in question for  $\ell_1$ -regularized linear regression (LASSO) and also for the more general well-studied procedure SLOPE. However, since it is not the focus of this article we do not explore the issue of bounding the complexity parameters involved beyond those two cases. Let us just mention that the question of finding data-dependent bounds on these quantities, or better still, computationally feasible data dependent bounds, is wide open.

### 3 Regularized risk minimization

Before we can properly describe the regularized tournament procedure, let us explain the path one may take when studying the standard regularized empirical risk minimization (RERM), which is the basis for the analysis of the regularized tournament procedure we introduce in what follows.

RERM selects a minimizer  $\hat{f}$  of the regularized empirical risk functional

$$\operatorname{argmin}_{f \in \mathcal{F}} \left( \frac{1}{N} \sum_{i=1}^N (f(X_i) - Y_i)^2 + \lambda \Psi(f) \right).$$

A typical bound on the performance of RERM for a well-chosen regularization parameter  $\lambda$  involves two estimates: first on the  $L_2$  distance  $\|\hat{f} - f^*\|_{L_2}$  and second, on the  $\Psi$ -distance  $\Psi(f - f^*)$ . The two estimates one can guarantee depend on structure of  $\mathcal{F}$ , the choice of  $\Psi$  and the tail behaviour of the functions involved.

Clearly, a minimizer of the regularized empirical risk functional also minimizes

$$\begin{aligned} & \frac{1}{N} \sum_{i=1}^N (f(X_i) - Y_i)^2 - (f^*(X_i) - Y_i)^2 + \lambda (\Psi(f) - \Psi(f^*)) \\ &= \frac{1}{N} \sum_{i=1}^N (f - f^*)^2(X_i) + \frac{2}{N} \sum_{i=1}^N (f^*(X_i) - Y_i)(f - f^*)(X_i) + \lambda (\Psi(f) - \Psi(f^*)) \\ &= Q_f + M_f + \lambda (\Psi(f) - \Psi(f^*)), \end{aligned} \tag{3.1}$$

where

$$Q_f = \frac{1}{N} \sum_{i=1}^N (f - f^*)^2(X_i)$$

is the quadratic component of the empirical excess squared risk and

$$M_f = \frac{2}{N} \sum_{i=1}^N (f^*(X_i) - Y_i)(f - f^*)(X_i)$$

is the corresponding “multiplier” component.

The key to the analysis of RERM is that the empirical minimizer  $\hat{f}$  satisfies that

$$Q_{\hat{f}} + M_{\hat{f}} + \lambda \left( \Psi(\hat{f}) - \Psi(f^*) \right) \leq 0 ,$$

because  $f^*$  is also a possible minimizer of (3.1). Hence, if one can show that for a large part of  $\mathcal{F}$ , (3.1) is positive, the empirical minimizer belongs to the complement of that set, which hopefully contains only functions that are ‘close’ to  $f^*$ .

Clearly, the quadratic component is always nonnegative. However, all the information one has on  $M_f$  is that  $\mathbb{E}M_f \geq 0$ , and that does not exclude the possibility that  $M_f$  is negative. Also, there is no off-hand reason why  $\lambda(\Psi(f) - \Psi(f^*))$  should be positive.

A possible way of controlling the performance of RERM has been suggested in [5]. The idea there is to find a wise choice of the radii  $\rho$  and  $r$  and a regularization parameter  $\lambda$ ; show that (3.1) is positive on the  $\Psi$ -sphere  $\{f \in \mathcal{F} : \Psi(f - f^*) = \rho\}$ , and therefore, by homogeneity properties of the estimates involved, that it is also positive outside the sphere. Finally, one may study the behaviour of (3.1) in the  $\Psi$ -ball  $\{f \in \mathcal{F} : \Psi(f - f^*) \leq \rho\}$ , which depends entirely on  $Q_f$  and  $M_f$ . Indeed, if  $\Psi(f - f^*) \leq \rho$  then

$$\lambda(\Psi(f) - \Psi(f^*)) \geq -\lambda\Psi(f - f^*) \geq -\lambda\rho$$

and the parameters  $r$  and  $\lambda$  are chosen in a way that ensures that if  $\Psi(f - f^*) \leq \rho$  and  $\|f - f^*\|_{L_2} \geq r$ , then  $Q_f$  dominates both  $M_f$  and  $\lambda\rho$ . Thus, the empirical excess risk functional is positive in

$$\{f \in \mathcal{F} : \Psi(f - f^*) \leq \rho, \quad \|f - f^*\|_{L_2} \geq r\} .$$

Combining all these observations, it follows that

$$\|\hat{f} - f^*\|_{L_2} \leq r \quad \text{and} \quad \Psi(\hat{f} - f^*) \leq \rho .$$

The heart of the argument (and the main role of the regularization term) is the fact that (3.1) is positive when  $\Psi(f - f^*) = \rho$  and  $\|f - f^*\|_{L_2} \leq r$ . Unfortunately, it is impossible to obtain a useful lower bound on  $Q_f$  in that region, and the only hope to ensure that (3.1) is positive is by showing that  $\lambda(\Psi(f) - \Psi(f^*))$  is large enough to defeat the potentially negative  $M_f$ . Indeed, in [5] the authors establish that for such functions, on the one hand,  $M_f \geq -(C/2)r$  and on the other, (non)-smoothness properties of  $\Psi$  imply that  $\lambda(\Psi(f) - \Psi(f^*)) \gtrsim \lambda\rho$ .

To illustrate what we mean by “non-smoothness properties of  $\Psi$ ”, let  $f \in \mathcal{F}$  satisfy  $\Psi(f - f^*) = \rho$  and  $\|f - f^*\|_{L_2} \leq r$ . Assume that there is a norm one (relative to  $\Psi$ ) linear functional  $z$  that is norming for both  $f - f^*$  and  $f^*$ , that is, if  $\Psi^*$  is the dual norm to  $\Psi$  then

$$\Psi^*(z) = 1, \quad z(f - f^*) = \Psi(f - f^*), \quad \text{and} \quad z(f^*) = \Psi(f^*) . \quad (3.2)$$

Since  $f^*$  and  $f - f^*$  have a common norming functional, it follows that

$$\Psi(f) - \Psi(f^*) \geq z(f) - z(f^*) = z(f - f^*) = \Psi(f - f^*) = \rho .$$

Hence,

$$M_f + \lambda(\Psi(f) - \Psi(f^*)) \geq -(C/2)r + \lambda\rho > 0 ,$$

provided that  $r$ ,  $\rho$  and  $\lambda$  are chosen accordingly.

Naturally, because all the functionals involved have to be norming for  $f^*$ ,  $\Psi$  cannot be smooth in  $f^*$  if we want (3.2) to hold for many functions in  $\mathcal{F}$ . Indeed, if it were smooth in  $f^*$ , such a norming functional would be unique. Therefore, if this type of argument is to be of any use,  $\partial\Psi_{f^*}$  (the subdifferential of  $\Psi$  in  $f^*$ ) has to be a large set.

To summarize, the results in [5] indicate that a good choice of  $\rho, r$  and  $\lambda$  satisfies the following:

- $f^*$  has enough ‘almost norming’ functionals. Specifically, enough to ensure that for any  $f \in \mathcal{F}$  for which  $\Psi(f - f^*) = \rho$  and  $\|f - f^*\|_{L_2} \leq r(\rho)$  there is a functional that is simultaneously almost norming for  $f^*$  and  $f - f^*$ . We make this somewhat vague description more explicit later.
- With high probability, if  $f \in \mathcal{B}_{f^*}(\rho)$  then

$$M_f \geq -\frac{C}{2} \max\{r^2, \|f - f^*\|_{L_2}^2\}$$

and

$$Q_f \geq C\|f - f^*\|_{L_2}^2 \quad \text{if} \quad \|f - f^*\|_{L_2} \geq r$$

for a suitable constant  $C$ .

- The choice of  $\rho$  and  $r(\rho)$  dictates the choice of  $\lambda$ . For example,

$$\frac{3C}{8} \cdot \frac{r^2(\rho)}{\rho} \leq \lambda \leq \frac{C}{2} \cdot \frac{r^2(\rho)}{\rho}$$

was shown in [5] to be a valid choice of  $\lambda$ .

The main observation in [5] is that if the above conditions are fulfilled, then RERM performed with a regularization parameter  $\lambda$  produces  $\hat{f}$  that satisfies

$$\Psi(\hat{f} - f^*) \leq \rho \quad \text{and} \quad \|\hat{f} - f^*\|_{L_2} \leq r .$$

Note that out of  $\rho, r$  and  $\lambda$ , only the latter is actually involved in the definition of the RERM procedure, while the other two serve as purely analytical instruments. Although there is no need to specify the ‘right’ values of  $\rho$  and  $r$ , these can be made explicit and are derived in [5]. In contrast, and unlike RERM, the regularized tournament procedure we introduce here requires the values of parameters like  $\rho$  and  $r$  as input, and the fact that these parameters may be specified is vital to its success. In return, what we gain over RERM is a procedure that is not damaged by the presence of heavy tails and performs well under considerably more general conditions.

## 4 The main result

Our main interest is in situations where the class  $\mathcal{F}$  is large. It is therefore natural to split it according to some notion of ‘simplicity’. The desired simplicity is expressed in form of a nested finite hierarchy of subsets of  $\mathcal{F}$ . More precisely, let  $\mathcal{F}_1, \dots, \mathcal{F}_K \subset \mathcal{F}$  be such that  $\mathcal{F} = \mathcal{F}_1 \supset \mathcal{F}_2 \supset \dots \supset \mathcal{F}_K$ . A function belonging to  $\mathcal{F}_\ell$  for a large value of  $\ell$  is considered as “simple”. For example, when dealing with sparse recovery problems in  $\mathbb{R}^n$ , a natural hierarchy in  $\mathbb{R}^d$  is determined according to the degree of sparsity of each element: the class  $\mathcal{F}_\ell$  consists of elements that are  $d/2^{\ell-1}$ -sparse relative to the standard basis.

Let  $\rho_1 > \rho_2 > \dots > \rho_K > 0$ , which serve a similar goal to that of  $\rho$  in RERM:  $\rho_\ell$  is used to identify the ‘right set’

$$\{f \in \mathcal{F} : \Psi(f - f^*) \leq \rho_\ell\} = \mathcal{F} \cap \mathcal{B}_{f^*}(\rho_\ell)$$

which is the set one should examine carefully if one believes that  $f^* \in \mathcal{F}_\ell$  (though obviously one does not have any knowledge of the identity of  $f^*$ ).

Given  $\rho_\ell$ , we consider the complexity of the localized class  $\mathcal{F} \cap \mathcal{B}_{f^*}(\rho_\ell)$ . More precisely, in order to make  $r_\ell$  independent of the unknown function  $f^*$ , we define

$$r_\ell = r^*(\mathcal{F}, \mathcal{F}_\ell, \rho_\ell) ,$$

and note that the belief behind each  $r_\ell$  is that  $f^* \in \mathcal{F}_\ell$ , hence the interest in centres that belong to  $\mathcal{F}_\ell$ .

The choice of  $r_\ell$  implies that it depends on  $\rho_\ell$ , and in a monotone fashion. However, we do not have total freedom in the choice of pairs  $(\rho_\ell, r_\ell)$ . As it happens, the correct choice of a pair  $\rho_\ell, r_\ell$  depends on the interplay between the hierarchy  $(\mathcal{F}_\ell)_{\ell=1}^K$  and the norm  $\Psi$  as we describe next. The core issue has been outlined in the previous section: if  $f^* \in \mathcal{F}_\ell$  then it must have enough norming functionals, and by “enough” we mean that for  $f \in \mathcal{F}$  that satisfies  $\Psi(f - f^*) = \rho_\ell$  and  $\|f - f^*\|_{L_2} \leq r_\ell$ , there is an almost norming functional of  $f^*$  that is also almost norming for  $f - f^*$ .

### 4.1 Properties of the hierarchy

Recall that  $\mathcal{F}$  is a subset of a normed space  $(E, \Psi)$ ;  $E$  is also a subspace of  $L_2(\mu)$ , though  $\Psi$  and  $\|\cdot\|_{L_2(\mu)}$  may have nothing to do with each other. Let  $B_{\Psi^*}$  and  $S_{\Psi^*}$  denote the unit ball and unit sphere in the dual space to  $(E, \Psi)$ , respectively. Therefore,  $B_{\Psi^*}$  consists of all the linear functionals  $z \in E^*$  for which  $\sup_{\{x \in E : \Psi(x)=1\}} |z(x)| \leq 1$ . A linear functional  $z^* \in S_{\Psi^*}$  is a norming functional for  $f \in E$  if  $z^*(f) = \Psi(f)$ .

**Definition 4.1.** Let  $\Gamma_f(\rho) \subset S_{\Psi^*}$  be the collection of functionals that are norming for some  $v \in \mathcal{B}_f(\rho/20)$ . Set

$$\Delta_\ell(\rho, r) = \inf_{f \in \mathcal{F}_\ell} \inf_h \sup_{z \in \Gamma_f(\rho)} z(h - f) ,$$

where the inner infimum is taken in the set

$$\{h \in \mathcal{F} : \Psi(h - f) = \rho \text{ and } \|h - f\|_{L_2} \leq r\} . \quad (4.1)$$

The idea behind the definition of  $\Delta_\ell(\rho, r)$  is the same as the one outlined earlier. It turns out that the main difficulty in the analysis of a regularized tournament is the behaviour of the empirical regularized excess risk in the set

$$\mathcal{F} \cap \{f : \Psi(f - f^*) = \rho_\ell\}.$$

Just like in the standard regularization framework, the regularization term  $\lambda(\Psi(h) - \Psi(f^*))$  saves the day when trying to deal with functions in  $h \in \mathcal{F}$  that satisfy  $\Psi(h - f^*) = \rho_\ell$  and  $\|h - f^*\|_{L_2} \leq r_\ell$ . It does so by ensuring that  $f^*$  (or functions close to  $f^*$ ) have enough norming functionals, and the ‘positive contribution’ of these norming functionals is measured by the parameter  $\Delta_\ell(\rho_\ell, r_\ell)$ .

Let us examine  $\Delta_\ell(\rho_\ell, r_\ell)$  and explain its meaning for some fixed values  $\rho, r > 0$ . Note that  $\Delta_\ell(\rho, r) \leq \rho$ . Indeed,  $\Gamma_f(\rho) \subset S_{\Psi^*}$  and if  $z \in S_{\Psi^*}$  and  $\Psi(h - f) \leq \rho$  then

$$|z(h - f)| \leq \Psi^*(z) \cdot \Psi(h - f) \leq \rho.$$

The interesting situation is when one can ensure a reverse inequality, that is, that  $\Delta_\ell(\rho, r)$  is proportional to  $\rho$ , say  $\Delta_\ell(\rho, r) \geq (4/5)\rho$ . Such a lower estimate on  $\Delta_\ell$  implies the following. Let  $f \in \mathcal{F}_\ell$  and  $h \in \mathcal{F}$  for which  $\Psi(h - f) = \rho$  and  $\|f - h\|_{L_2} \leq r$ . It follows that there is some  $z \in S_{\Psi^*}$  and  $v \in \mathcal{B}_f(\rho/20)$  such that  $z$  is norming for  $v$  and  $z(h) - z(f) \geq \Delta_\ell(\rho, r)$ . Therefore,

$$\begin{aligned} \Psi(h) - \Psi(f) &= \Psi(h) - \Psi(v + (f - v)) \geq \Psi(h) - \Psi(v) - \Psi(f - v) \\ &\geq z(h) - z(v) - \Psi(f - v) \geq z(h) - z(f) - 2\Psi(f - v) \\ &\geq \Delta_\ell(\rho, r) - \rho/10 \geq 3\rho/5, \end{aligned}$$

which is precisely the type of lower bound we require for the regularized functional.

Obviously, ensuring that  $\Delta_\ell(\rho, r) \geq (4/5)\rho$  becomes simpler when the set  $\Gamma_f(\rho)$  is large. In the extreme case, when  $\rho > 30\Psi(f)$ , it follows that  $\mathcal{B}_f(\rho/20)$  contains a nontrivial  $\Psi$ -ball around 0; thus,  $\Gamma_f(\rho) = S_{\Psi^*}$  and  $\Delta_\ell(\rho, r) = \rho$ . The other extreme is if  $\rho$  is very small and one is left only with the functionals that are norming for  $f$  itself.

Intuitively, the right choice of  $\rho_\ell$  is the smallest one for which, for  $r_\ell = r^*(\mathcal{F}, \mathcal{F}_\ell, \rho_\ell)$ , one has  $\Delta_\ell(\rho_\ell, r_\ell) \geq 4\rho_\ell/5$ . With that in mind, let us define a ‘good hierarchy’ of  $\mathcal{F}$ :

**Definition 4.2.** *The sequence  $(\mathcal{F}_\ell, \rho_\ell)_{\ell=1}^K$  is compatible if*

- (1)  $\mathcal{F} = \mathcal{F}_1 \supset \mathcal{F}_2 \supset \dots \supset \mathcal{F}_K$  is a finite hierarchy;
- (2)  $(\rho_\ell)_{\ell=1}^K$  is decreasing and  $r_\ell = r^*(\mathcal{F}, \mathcal{F}_\ell, \rho_\ell)$ ;
- (3) for every  $1 \leq \ell \leq K$ ,  $\Delta_\ell(\rho_\ell, r_\ell) \geq 4\rho_\ell/5$ .

We allow the choice of  $\rho_\ell = r_\ell = \infty$  for  $\ell = 1, \dots, \ell_0$ . In such cases the compatibility condition is to be verified from  $\ell_0 + 1$  onward.

With all the definitions set in place, let us formulate the main result of this article.

**Theorem 4.3.** *Let  $\mathcal{F}$  and  $(X, Y)$  satisfy Assumption 2.1. Consider a hierarchy  $(\mathcal{F}_\ell)_{\ell=1}^K$  and set  $\ell^*$  to be the largest  $\ell$  for which  $f^* \in \mathcal{F}_\ell$ . Assume that  $(\mathcal{F}_\ell, \rho_\ell)_{\ell=1}^K$  is compatible, set  $r_\ell = r^*(\mathcal{F}, \mathcal{F}_\ell, \rho_\ell)$  and consider a decreasing sequence  $(\hat{r}_\ell)_{\ell=1}^K$  that satisfies  $\hat{r}_\ell \geq r_\ell$ .*

There is a regularized procedure that receives as input the values  $(\rho_\ell)_{\ell=1}^K$  and  $(\hat{r}_\ell)_{\ell=1}^K$ , and a sample  $(X_i, Y_i)_{i=1}^{3N}$ . It returns as output a function  $\hat{h} \in \mathcal{F}$ , for which, with probability at least  $1 - 2 \sum_{\ell=1}^{\ell^*} \exp(-c_0(L)N \min\{1, \sigma^{-2}\hat{r}_\ell^2\})$ ,

$$\Psi(\hat{h} - f^*) \leq \rho_{\ell^*}, \quad \|\hat{h} - f^*\|_{L_2} \leq c_1(L)\hat{r}_{\ell^*}, \quad \text{and} \quad R(\hat{h}) - R(f^*) \leq c_2(L)\hat{r}_{\ell^*}^2,$$

for constants  $c_0, c_1$  and  $c_2$  that depend only on  $L$  from Assumption 2.1.

In other words, the regularized procedure yields (almost) the optimal accuracy-confidence tradeoff for any  $r \geq r_{\ell^*}$ , and as such, it behaves as if it “saw” the location of  $f^*$  in the hierarchy without actually knowing it. Moreover, the best accuracy one can attain from Theorem 4.3, namely,  $r_{\ell^*}$ , is essentially the best known error rate of any learning procedure taking values only in  $\mathcal{F}_{\ell^*}$ .

## 5 The components of the procedure

The main contribution of this paper is the novel learning procedure, announced in Theorem 4.3. This “regularized tournament procedure” is based on the median-of-means tournament, introduced in [8], that was shown to achieve optimal accuracy/confidence tradeoff in a convex class under general conditions. Just like the median-of-means tournament, the regularized version has three components: the “distance oracle”; the “elimination phase”; and the “champions league”. The procedure is actually performed  $K$  times in  $\mathcal{F}$ , with the  $\ell$ -th stage based on the belief that  $f^* \in \mathcal{F}_\ell$ . This belief results in  $K$  different distance oracles, elimination phases and champions leagues. Still, all  $K$  tournaments use the same sample  $(X_i, Y_i)_{i=1}^{3N}$ .

### The $\ell$ -distance oracle

In each stage  $\ell = 1, \dots, K$  of the procedure, one initially uses a modification of the distance oracle from [8].

The  $\ell$ -distance oracle is a data-dependent procedure that provides information on the distances between functions. It is used for any pair  $f, h \in \mathcal{F}$ , and aims at determining if  $\Psi(f - h) \geq \rho_\ell$ , or, if  $\Psi(f - h) \leq \rho_\ell$ , whether  $\|f - h\|_{L_2} \geq \hat{r}_\ell$ . Note that  $\Psi$  is a known norm and therefore,  $\Psi(f - h)$  is known for any pair  $f, h \in \mathcal{F}$  but  $\|f - h\|_{L_2}$  needs to be (crudely) estimated.

We work under Assumption 2.1. Fix a positive integer  $n \leq N$  whose value is an appropriately chosen constant that only depends on  $L$ . Assume without loss of generality that  $n$  divides  $N$  and partition  $\{1, \dots, N\}$  into  $n$  disjoint blocks  $(I_j)_{j=1}^n$ , each one of cardinality  $m = N/n$ .

Given  $w \in \mathbb{R}^N$ , set

$$v_j(w) = \frac{1}{m} \sum_{i \in I_j} w_i$$

and let  $\text{Med}_m(w)$  be a median of  $\{v_1, \dots, v_n\}$ . We call  $\text{Med}_m(w)$  the *median of means* of the vector  $w$ .

For a sample  $\mathcal{C}_1 = (X_i)_{i=1}^N$  and functions  $f$  and  $h$ , let  $w = (|f(X_i) - h(X_i)|)_{i=1}^N$  and set

$$\Phi_{\mathcal{C}_1}(f, h) = \text{Med}_m(w) .$$

The behaviour of  $\Phi$  described below has been established in Mendelson [11] (see also Lugosi and Mendelson [8]):



**Proposition 5.1.** *Let  $\mathcal{F}$  satisfy Assumption 2.1. There exist constants  $\kappa, \eta, c > 0$  and  $0 < \alpha < 1 < \beta$ , all of them depending only on  $L$  for which the following holds. Set  $1 \leq \ell \leq K$  and  $\rho > 0$ , let  $r > r^*(\mathcal{F}, \mathcal{F}_\ell, \rho)$  ( $r^*$  is defined relative to the constants  $\kappa$  and  $\eta$ ) and fix  $f \in \mathcal{F}_\ell$ . Then, with probability at least  $1 - 2\exp(-cN)$ , for any  $h \in \mathcal{F}$  that satisfies  $\Psi(f, h) \leq \rho$ ,*

- *if  $\Phi_{\mathcal{C}_1}(f, h) \geq \beta r$  then  $\beta^{-1}\Phi_{\mathcal{C}_1}(f, h) \leq \|f - h\|_{L_2} \leq \alpha^{-1}\Phi_{\mathcal{C}_1}(f, h)$ , and*
- *if  $\Phi_{\mathcal{C}_1}(f, h) < \beta r$  then  $\|f - h\|_{L_2} \leq (\beta/\alpha)r$ .*

The proof of Proposition 5.1 is a direct outcome of Proposition 3.2 from [8], applied to the set  $\mathcal{F} \cap \mathcal{B}_f(\rho)$  for a fixed centre  $f \in \mathcal{F}_\ell$ .

Based on Proposition 5.1 we may introduce the  $\ell$ -distance oracle  $\mathcal{DO}_\ell$  as follows:

**Definition 5.2.** *let  $\hat{r}_\ell > r_\ell = r^*(\mathcal{F}, \mathcal{F}_\ell, \rho_\ell)$ . Let  $f_1, f_2 \in \mathcal{F}$ . If  $\Psi(f_1 - f_2) > \rho_\ell$  or if  $\Psi(f_1 - f_2) \leq \rho_\ell$  and  $\Phi_{\mathcal{C}_N}(f_1, f_2) \geq \beta\hat{r}_\ell$ , set  $\mathcal{DO}_\ell(f_1, f_2) = 1$ ; otherwise set  $\mathcal{DO}_\ell(f_1, f_2) = 0$ .*

Thanks to Proposition 5.1, it follows that for a fixed centre  $f \in \mathcal{F}_\ell$  (which is selected as  $f^*$  in what follows) with probability at least  $1 - 2\exp(-cN)$  if  $h \in \mathcal{F}$ , and  $\mathcal{DO}_\ell(f, h) = 0$  then  $\Psi(h - f) \leq \rho_\ell$  and  $\|f - h\|_{L_2} \leq \hat{r}_\ell$ .

**Remark 5.3.** *Although Proposition 5.1 is formulated for a designated single centre  $f$ , it is straightforward to extend it to any centre in  $\mathcal{F}$  and obtain a uniform distance oracle that holds for any pair  $f, h \in \mathcal{F}$ .*

After one obtains enough information of distances between functions in  $\mathcal{F}$ , the rest of the  $\ell$ -th stage of the tournament consists of two rounds: a regularized elimination phase—which we describe first—, followed by a champions league round.

### $\ell$ -elimination phase

Fix  $1 \leq \ell \leq K$  and let  $f, h \in \mathcal{F}$ . A regularized match between  $f$  and  $h$  is defined as follows. First, the  $\ell$ -distance oracle defined above uses the first part of the sample  $\mathcal{C}_1 = (X_i, Y_i)_{i=1}^N$  to determine the value of  $\mathcal{DO}_\ell(f, h)$ . If  $\mathcal{DO}_\ell(f, h) = 0$ , the match is abandoned.

Each match that is allowed to take place by the  $\ell$ -distance oracle is played using the second part of the sample,  $(X_i, Y_i)_{i=N+1}^{2N}$ . The sub-sample is partitioned to  $n$  blocks  $(I_j)_{j=1}^n$  of cardinality  $m = N/n$  where  $n$  is chosen as  $\theta_1(L)N \min\{1, \hat{r}_\ell^2/\sigma_4^2\}$ . We set

$$\lambda_\ell = \theta_2(L) \frac{\hat{r}_\ell^2}{\rho_\ell},$$

with the choices of both constants  $\theta_1$  and  $\theta_2$  specified below.

**Definition 5.4.** *The function  $f$  defeats  $h$  if*

$$\frac{1}{m} \sum_{i \in I_j} ((h(X_i) - Y_i)^2 - (f(X_i) - Y_i)^2) + \lambda_\ell(\Psi(h) - \Psi(f)) > 0$$

*on a majority of the blocks  $I_j$ .*

The set of winners  $\mathcal{H}'_\ell$  of the  $\ell$ -th elimination round consists of all the functions in  $\mathcal{F}$  that have not lost a single match.

The role of the elimination round is to ‘exclude’ functions that are far from  $f^*$  (without knowing the identity of  $f^*$ , of course). To that end, it suffices to show that, with high probability,  $f^*$  wins all the matches it takes part in. Indeed, that implies that matches between  $f^*$  and any  $h \in \mathcal{H}'_\ell$  must have been abandoned, and therefore,  $\mathcal{DO}_\ell(f^*, h) = 0$ , that is,

$$\Psi(f^* - h) \leq \rho_\ell \quad \text{and} \quad \|f^* - h\|_{L_2} \leq (\beta/\alpha)\hat{r}_\ell. \quad (5.1)$$

The next theorem describes the outcome of the elimination phase. Its proof may be found in Section 5.1.

**Proposition 5.5.** *Using the notation above, if  $f^* \in \mathcal{F}_\ell$  then, with probability at least*

$$1 - 2 \exp(-c_0 N \min\{1, \sigma_4^{-2} \hat{r}_\ell^2\}) ,$$

*$f^*$  wins all the matches it participates in. In particular, on that event, if  $h \in \mathcal{H}'_\ell$ , then (5.1) holds.*

### $\ell$ -champions league

Once Proposition 5.5 is established, the second part of  $\ell$ -th stage of the tournament is the selection of a set of “winners”. In order to do that, we run the same champions league tournament used in [8], performed in each one of the sets  $\mathcal{H}'_\ell$ . The crucial point is that if  $f^* \in \mathcal{F}_\ell$  then  $\mathcal{H}'_\ell$  satisfies the necessary conditions for a champions league tournament: that  $f^* \in \mathcal{H}'_\ell$  and all the functions  $h \in \mathcal{H}'_\ell$  have a mean-squared error at most  $\sim \hat{r}_\ell$ . In what follows we consider such sets  $\mathcal{H}'_\ell$ .

The  $\ell$ -champions league consists of matches that use a third part of the sample  $(X_i, Y_i)_{i=2N+1}^{3N}$ . Let  $(I_j)_{j=1}^n$  be the partition of  $\{2N+1, \dots, 3N\}$  to  $n$  blocks, for the same value of  $n$  as in the  $\ell$ -elimination phase.

The matches in the champions league consist of “home-and-away” legs.

**Definition 5.6.** *Let  $\alpha$  and  $\beta$  be as in Proposition 5.1, set  $c = \beta/\alpha$  and for  $f, h \in \mathcal{H}'_\ell$ , let  $\Psi_{h,f} = (h(X) - f(X))(f(X) - Y)$ . The function  $f$  wins its home match against  $h$  if*

$$\frac{2}{m} \sum_{i \in I_j} \Psi_{h,f}(X_i, Y_i) \geq -(2c\hat{r}_\ell)^2/10$$

*on more than  $n/2$  of the blocks  $I_j$ .*

*The set of winners  $\mathcal{H}_\ell$  as the set of all the “champions” in  $\mathcal{H}'_\ell$  that win all of their home matches.*

The outcome of the  $\ell$ -champions league is as follows:

**Proposition 5.7.** *Let  $\mathcal{H}'_\ell$  as above. With probability at least*

$$1 - 2 \exp(-c_0 N \min\{1, \sigma_4^{-2} \hat{r}_\ell^2\})$$

*with respect to  $(X_i, Y_i)_{i=2N+1}^{3N}$ , the set of winners  $\mathcal{H}_\ell$  contains  $f^*$ , and if  $h \in \mathcal{H}_\ell$  then*

$$R(h) - R(f^*) \leq 16c^2 \hat{r}_\ell^2 .$$

Proposition 5.7 is an immediate outcome of Proposition 3.8 from [8] for  $H = \mathcal{H}'_\ell$  and using the fact that  $f^* \in \mathcal{H}'_\ell$  and that if  $h \in \mathcal{H}'_\ell$  then  $\|h - f^*\|_{L_2} \leq (\beta/\alpha)\widehat{r}_\ell$ .

Combining all these observations, Corollary 5.8 describes the outcome of the  $\ell$ -tournament procedure.

**Corollary 5.8.** *Using the above notation, for any  $1 \leq \ell \leq K$  the following holds. If  $f^* \in \mathcal{F}_\ell$ , then with probability at least*

$$1 - 2 \exp(-c_1(L)N \min\{1, \sigma_4^{-2}\widehat{r}_\ell^2\})$$

*with respect to  $(X_i, Y_i)_{i=1}^{3N}$ , the set of winners  $\mathcal{H}_\ell$  satisfies:*

- $f^* \in \mathcal{H}_\ell$ , and
- for any  $h \in \mathcal{H}_\ell$ ,

$$\Psi(h - f^*) \leq \rho_\ell, \quad \|h - f^*\|_{L_2} \leq (\beta/\alpha)\widehat{r}_\ell, \quad \text{and} \quad R(h) - R(f^*) \leq 16(\beta/\alpha)^2 \widehat{r}_\ell^2.$$

### Selection of a final winner

Once the regularized elimination phase and the regularized champions league has been executed for all  $\ell = 1, \dots, K$  stages, we have  $K$  sets  $\mathcal{H}_\ell$  consisting of stage winners. Some of these sets may be empty but, as we have seen it above, for those indices  $\ell$  for which  $f^* \in \mathcal{F}_\ell$ , with high probability the set  $\mathcal{H}_\ell$  contains  $f^*$ . In order to select the final “winner”, let  $\ell_1$  be the largest integer  $1 \leq \ell \leq K$  for which  $\bigcap_{j \leq \ell} \mathcal{H}_j \neq \emptyset$ . The procedure returns any  $\widehat{h} \in \bigcap_{j \leq \ell_1} \mathcal{H}_j$ .

Clearly, on a high probability event,  $\ell_1 \geq \ell^*$ . Therefore, the selected function  $\widehat{h}$  belongs to  $\mathcal{H}_{\ell^*}$ . Moreover, recalling that  $(\widehat{r}_\ell)_{\ell=1}^K$  is decreasing,

$$\Psi(\widehat{h} - f^*) \leq \widehat{r}_{\ell^*}, \quad \|\widehat{h} - f^*\|_{L_2} \leq (\beta/\alpha)\widehat{r}_{\ell^*} \quad \text{and} \quad R(\widehat{h}) - R(f^*) \leq c(\beta/\alpha)^2 \widehat{r}_{\ell^*}^2,$$

which completes the proof of Theorem 4.3. ■

## 5.1 Proof of Proposition 5.5—highlights

The main difference between the regularized tournament and the “unregularized” one from [8] is the regularized elimination phase. To explain why this elimination phase preforms well even when  $\mathcal{F}$  is very large, define, for each block  $I_j$  ( $j = 1, \dots, n$ ),

$$B_{h,f}^\lambda(j) = \frac{1}{m} \sum_{i \in I_j} ((h(X_i) - Y_i)^2 - (f(X_i) - Y_i)^2) + \lambda(\Psi(h) - \Psi(f)). \quad (5.2)$$

Note that the regularized empirical excess risk of  $h$  on block  $I_j$  is  $B_{h,f^*}^\lambda(j)$ .

Consider the  $\ell$ -th stage of the regularized tournament. The assertion of Proposition 5.5 is that, if  $f^* \in \mathcal{F}_\ell$ , then it is a winner of all the elimination phase matches it participates in. Hence, Proposition 5.5 is proved once we ensure that for the right choice of  $\lambda = \lambda_\ell$ , with high probability, if  $h \in \mathcal{F}$  and  $\mathcal{DO}_\ell(f^*, h) = 1$  then  $B_{h,f^*}^\lambda(j)$  is positive for most of the blocks  $I_j$ .

To that end, observe that

$$\begin{aligned} & \frac{1}{m} \sum_{i \in I_j} ((h(X_i) - Y_i)^2 - (f^*(X_i) - Y_i)^2) \\ &= \frac{1}{m} \sum_{i \in I_j} (h - f^*)^2(X_i) + \frac{2}{m} \sum_{i \in I_j} (h - f^*)(X_i) \cdot (f^*(X_i) - Y_i), \end{aligned}$$

which is the natural decomposition of the empirical excess risk functional into its quadratic and multiplier components. Setting

$$\mathbb{Q}_{h,f}(j) = \frac{1}{m} \sum_{i \in I_j} (h - f)^2(X_i) \quad \text{and} \quad \mathbb{M}_{h,f}(j) = \frac{2}{m} \sum_{i \in I_j} (h - f)(X_i) \cdot (f(X_i) - Y_i),$$

we have

$$\begin{aligned} B_{h,f^*}^\lambda(j) &= \frac{1}{m} \sum_{i \in I_j} ((h(X_i) - Y_i)^2 - (f^*(X_i) - Y_i)^2) + \lambda(\Psi(h) - \Psi(f^*)) \\ &= \mathbb{Q}_{h,f^*}(j) + \mathbb{M}_{h,f^*} + \lambda(\Psi(h) - \Psi(f^*)) . \end{aligned}$$

The first observation we require is a version of a deterministic result from [5, Theorem 3.2] (see the appendix for the proof).

**Lemma 5.9.** *Let  $f^* \in \mathcal{F}_\ell$  and  $h \in \mathcal{F}$  for which either  $\Psi(h - f^*) = \rho$ , or  $\Psi(h - f^*) < \rho$  and  $\|h - f^*\|_{L_2} \geq r$ . Assume that  $\Delta_\ell(\rho, r) \geq 4\rho/5$ , that  $\lambda$  satisfies*

$$\frac{C}{2} \cdot \frac{r^2}{\rho} \leq \lambda \leq \frac{3C}{4} \cdot \frac{r^2}{\rho}. \quad (5.3)$$

Assume further that

$$\mathbb{M}_{h,f^*}(j) \geq -(C/4) \max \{ \|h - f^*\|_{L_2}^2, r^2 \}, \quad (5.4)$$

and if also  $\|h - f^*\|_{L_2} \geq r$  then

$$\mathbb{Q}_{h,f^*}(j) \geq C \|h - f^*\|_{L_2}^2. \quad (5.5)$$

Then

$$\mathbb{Q}_{h,f^*}(j) + \mathbb{M}_{h,f^*}(j) + \lambda(\Psi(h) - \Psi(f^*)) > 0 .$$

Thanks to Lemma 5.9, all that is required to prove Proposition 5.5 is to verify that with the requested probability, (5.4) and (5.5) hold uniformly in  $h$  and on a majority of the blocks  $I_j$ , provided that  $f^* \in \mathcal{F}_\ell$ . Indeed, we have the following lemma, proved in the appendix:

**Lemma 5.10.** *There exists an absolute constant  $c$  and a constant  $C_1 = C_1(L, \tau)$  for which the following holds. Let  $f^* \in \mathcal{F}_\ell$ . For  $0 < \tau < 1$ , with probability at least  $1 - 2\exp(-c\tau^2 n)$ , for every  $h \in \mathcal{B}_{f^*}(\rho)$  that satisfies  $\|f - f^*\|_{L_2} \geq \hat{r}_\ell$ , we have*

$$|\{j : \mathbb{Q}_{f,f^*}(j) \geq C_1 \|f - f^*\|_{L_2}^2\}| \geq (1 - \tau) n$$

and

$$\left| \left\{ j : \mathbb{M}_{f,f^*}(j) \leq -\frac{C_1}{4} \|f - f^*\|_{L_2}^2 \right\} \right| \leq \tau n .$$

Moreover, for every  $h \in \mathcal{B}_{f^*}(\rho)$  that satisfies  $\|f - f^*\|_{L_2} \leq \hat{r}_\ell$ , we have

$$\left| \left\{ j : \mathbb{M}_{f,f^*}(j) \leq -\frac{C_1}{4} r^2 \right\} \right| \leq \tau n .$$

It follows from Lemma 5.10 that if  $\tau < 1/4$  then with probability at least  $1 - 2\exp(-c\tau^2 n)$ , for every  $h$  as in Lemma 5.9, conditions (5.4) and (5.5) hold for  $C = C_1$  and  $r = \hat{r}_\ell$  on the majority of the blocks  $I_j$ . Hence, setting  $\tau = 1/10$ , on an event with probability at least  $1 - 2\exp(-cn)$ ,  $f^*$  wins all the matches it participates in and that are allowed to take place by the  $\ell$ -distance oracle, as we require.

## 6 Examples

Let us return to the two examples, LASSO and SLOPE, and to the proofs of Theorems 1.3 and 1.4.

Recall that the assumptions here are stronger than Assumption 2.1. Namely,  $X$  is assumed to be isotropic, and for every  $t \in \mathbb{R}^d$  and  $1 \leq p \leq c \log d$ ,

$$\| \langle t, X \rangle \|_{L_p} \leq C \sqrt{p} \| \langle t, X \rangle \|_{L_2} = C \sqrt{p} \| t \|_2 . \quad (6.1)$$

In other words, linear forms satisfy a sub-Gaussian moment growth, but only up to a rather low exponent—logarithmic in the dimension of the underlying space. This moment assumption is a sufficient and almost necessary condition for the celebrated basis pursuit procedure to have a unique minimizer (see [6]), and as such, it is a natural assumption when studying sparsity-driven bounds. Note that under such a moment assumption, combined with the only condition that  $\xi \in L_4$ , a sub-Gaussian tail estimate for the supremum

$$\sup_{t \in T} \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N \xi_i \langle t, X_i \rangle - \mathbb{E} \xi \langle t, X \rangle \right| ,$$

is totally out of question even when  $|T| = 1$ .

Thankfully, (6.1) suffices to obtain bounds on the expectation of empirical and multiplier processes, as long as the indexing set has enough symmetries.

**Definition 6.1.** *Given a vector  $x = (x_i)_{i=1}^n$ , let  $(x_i^*)_{i=1}^n$  be the non-increasing rearrangement of  $(|x_i|)_{i=1}^n$ .*

*The normed space  $(\mathbb{R}^d, \| \cdot \|)$  is  $K$ -unconditional with respect to the basis  $\{e_1, \dots, e_d\}$  if for every  $x \in \mathbb{R}^d$  and every permutation of  $\{1, \dots, n\}$ ,*

$$\left\| \sum_{i=1}^d x_i e_i \right\| \leq K \left\| \sum_{i=1}^d x_{\pi(i)} e_i \right\| ,$$

*and if  $y \in \mathbb{R}^d$  and  $x_i^* \leq y_i^*$  for  $1 \leq i \leq d$  then*

$$\left\| \sum_{i=1}^d x_i e_i \right\| \leq K \left\| \sum_{i=1}^d y_i e_i \right\| .$$

There are many natural examples of  $K$ -unconditional spaces, most notably, all the  $\ell_p$  spaces. Moreover, the norm  $\|z\| = \sup_{v \in V} \sum_{i=1}^n v_i^* z_i^*$  is 1-unconditional. In fact, if  $V \subset \mathbb{R}^n$  is closed under coordinate permutations and reflections (sign-changes), then  $\| \cdot \| = \sup_{v \in V} | \langle \cdot, v \rangle |$  is 1-unconditional in the sense of Definition 6.1.

The following fact has been recently established in [12]:

**Theorem 6.2.** *There exists an absolute constant  $c_1$  and for  $K \geq 1$ ,  $L \geq 1$  and  $q_0 > 2$  there exists a constant  $c_2$  that depends only on  $K$ ,  $L$  and  $q_0$  for which the following holds. Consider*

- $V \subset \mathbb{R}^d$  for which the norm  $\| \cdot \| = \sup_{v \in V} | \langle v, \cdot \rangle |$  is  $K$ -unconditional with respect to the basis  $\{e_1, \dots, e_d\}$ ,
- $\xi \in L_{q_0}$  for some  $q_0 > 2$ ,

- an isotropic random vector  $X \in \mathbb{R}^d$  that satisfies

$$\max_{1 \leq j \leq d} \sup_{1 \leq p \leq c_1 \log d} \frac{\|\langle X, e_j \rangle\|_{L_p}}{\sqrt{p}} \leq L .$$

If  $(X_i, \xi_i)_{i=1}^N$  are independent copies of  $(X, \xi)$  then

$$\mathbb{E} \sup_{v \in V} \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N (\xi_i \langle X_i, v \rangle - \mathbb{E} \xi \langle X, v \rangle) \right| \leq c_2 \|\xi\|_{L_{q_0}} \ell_*(V) ,$$

where  $\ell_*(V) = \mathbb{E} \sup_{v \in V} \sum_{i=1}^d g_i v_i$  and  $G = (g_i)_{i=1}^d$  is a standard Gaussian vector in  $\mathbb{R}^d$ .

Therefore, as long as  $V$  is sufficiently symmetric and linear forms exhibit a sub-Gaussian moment growth up to  $p \sim \log d$ , the expectations of empirical and multiplier processes indexed by  $V$  behave as if  $X$  were the standard Gaussian vector and  $\xi$  were independent of  $X$ . In the cases we are interested in the indexing sets have enough symmetries, and since  $\xi \in L_4$ , the conditions of Theorem 6.2 hold for  $q_0 = 4$ .

## 6.1 The tournament LASSO

As we noted previously, the regularization function for LASSO is  $\Psi(t) = \|t\|_1$ , and for any  $h = \langle t_0, \cdot \rangle$  and  $\rho, r > 0$ ,

$$\mathcal{F}_{h,\rho,r} = \{\langle t, \cdot \rangle \in \mathbb{R}^d : t \in \rho B_1^d \cap r B_2^d\} .$$

Hence, for a fixed radius  $\rho$  the parameters  $r_E$  and  $\bar{r}_M$  are defined using the fixed-point conditions

$$\mathbb{E} \sup_{t \in \rho B_1^d \cap r B_2^d} \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N \varepsilon_i \langle t, X_i \rangle \right| \leq \kappa \sqrt{N} r \quad (6.2)$$

and

$$\mathbb{E} \sup_{t \in \rho B_1^d \cap r B_2^d} \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N \varepsilon_i \xi_i \langle t, X_i \rangle \right| \leq \kappa \sqrt{N} r^2 \quad (6.3)$$

respectively. Note that the indexing set  $V_{\rho,r} = \rho B_1^d \cap r B_2^d$  is invariant under coordinate permutations and sign reflections, and therefore satisfies the conditions of Theorem 6.2. Hence, an upper bound on  $r_E$  follows if

$$\mathbb{E} \sup_{t \in \rho B_1^d \cap r B_2^d} \sum_{i=1}^N g_i t_i \leq \kappa \sqrt{N} r , \quad (6.4)$$

while for an upper estimate on  $\bar{r}_M$  it suffices to ensure that

$$\|\xi\|_{L_4} \mathbb{E} \sup_{t \in \rho B_1^d \cap r B_2^d} \sum_{i=1}^N g_i t_i \leq \kappa \sqrt{N} r^2 . \quad (6.5)$$

Observe that both estimates cannot be improved. They are tight bounds on (6.2) and (6.3) when, for example,  $X = (g_1, \dots, g_d)$  and  $\xi$  is a Gaussian variable that is independent of  $X$ .

The added value in (6.4) and (6.5) is that if  $r$  satisfies these inequalities then, necessarily,  $\max\{\lambda_{\mathbb{Q}}, \lambda_{\mathbb{M}}\} \leq r$  (for a well chosen constant  $\kappa$ ). This is an immediate consequence of Sudakov's inequality, which implies that for some absolute constant  $c > 0$ , for any  $T \subset \mathbb{R}^d$  and any  $\varepsilon > 0$ ,

$$\varepsilon \sqrt{\log \mathcal{M}(T, \varepsilon B_2^d)} \leq c \mathbb{E} \sup_{t \in T} \sum_{i=1}^d g_i t_i \equiv \ell_*(T) .$$

Thus, when applied to the definition on  $\lambda_{\mathbb{Q}}$  one obtains

$$\log \mathcal{M}(\rho B_1^d \cap r B_2^d, \eta r B_2^d) \leq \frac{\ell_*^2(\rho B_1^d \cap r B_2^d)}{(\eta r)^2} \leq \kappa^2 N ,$$

that is, it suffices that

$$\ell_*(\rho B_1^d \cap r B_2^d) \leq \eta \kappa \sqrt{N} r ,$$

which is precisely the type of condition in (6.4).

We now come to the question of selecting the parameters  $\rho_\ell$ ,  $r_\ell$  and  $\lambda_\ell$  for the tournament LASSO. The choice requires several observations that have been established in [5].

First, the requirement that  $\Delta(\rho, r) \geq 4\rho/5$  forces some constraint on the choice of  $\rho$  and  $r$ . To simplify things, assume that  $t^* = \operatorname{argmin}_{t \in \mathbb{R}^d} \mathbb{E}(\langle X, t \rangle - Y)^2$  is supported on  $I \subset \{1, \dots, d\}$  and that  $|I| \leq s$ . Recall that by the definition of  $\Delta(\rho, r)$ , the fact that  $\Psi(t) = \|t\|_1$  and since  $X$  is isotropic, it suffices to consider vectors  $t \in \mathbb{R}^d$  for which  $\|t - t^*\|_1 = \rho$  and  $\|t - t^*\|_2 \leq r$ .

For such  $t$ ,

$$\|t\|_1 - \|t^*\|_1 = \sum_{i \in I^c} |t_i| + \sum_{i \in I} (|t_i| - |t_i^*|) \geq \sum_{i \in I^c} |t_i| - \sum_{i \in I} |t_i - t_i^*| ,$$

and since  $|I| \leq s$ ,

$$\sum_{i \in I} |t_i - t_i^*| \leq \sqrt{|I|} \|t - t^*\|_2 \leq \sqrt{s} r .$$

Therefore,

$$\sum_{i \in I^c} |t_i| = \sum_{i \in I^c} |t_i - t_i^*| = \sum_{i=1}^n |t_i - t_i^*| - \sum_{i \in I} |t_i - t_i^*| \geq \rho - \sqrt{s} r .$$

On the other hand, there is a functional  $z$  that is norming for both  $t^*$  and  $P_{I^c} t = \sum_{i \in I^c} t_i e_i$ ; hence,

$$\begin{aligned} z(t - t^*) &\geq z(P_{I^c}(t - t^*)) - \sum_{i \in I} |t_i - t_i^*| = \sum_{i \in I^c} |t_i - t_i^*| - \sum_{i \in I} |t_i - t_i^*| \\ &\geq \rho - 2\sqrt{s} r \geq \frac{4\rho}{5} \end{aligned}$$

as long as  $s \lesssim (\rho/r)^2$ . This shows that, as long as the ratio  $\rho/r$  is larger than the square-root of the degree of sparsity of vectors we are interested in,  $\Delta(\rho, r) \geq (4/5)\rho$  as our procedure requires. A similar observation is true if  $t^*$  is not sparse, but rather well approximated by an  $s$ -sparse vector (see [5] for a detailed argument).

Set  $k = (\rho/r)^2$  and assume without loss of generality that  $k$  is an integer. We also restrict ourselves to values  $1 \leq k \leq d$ , intuitively because the above implies that  $(\rho/r)^2$  should capture the degree of sparsity. Recall that

$$\ell_*(\rho B_1^d \cap r B_2^d) = r \ell_*(\sqrt{k} B_1^d \cap B_2^d) \leq C r \sqrt{k \log(ed/k)} = C \rho \sqrt{\log(edr^2/\rho^2)}$$



(see, e.g. [5] for the standard proof). Hence, (6.4) becomes

$$C\rho\sqrt{\log(edr^2/\rho^2)} \leq \kappa\sqrt{N}r, \quad (6.6)$$

while (6.5) implies

$$\|\xi\|_{L_4} \cdot C\rho\sqrt{\log(edr^2/\rho^2)} \leq \kappa\sqrt{N}r^2. \quad (6.7)$$

We consider only the case  $N \leq Cd$ , which is the more interesting range in sparse recovery—when the number of given linear measurements is significantly smaller than the dimension of the underlying space. An argument following the same path may be used when  $N \geq Cd$  and we omit it.

It follows from a rather tedious computation that (6.6) holds provided that

$$r \geq c \frac{\rho}{\kappa\sqrt{N}} \sqrt{\log\left(\frac{cd}{\kappa N}\right)}, \quad (6.8)$$

and it follows from (6.7) that

$$r^2 \geq c\rho \frac{\|\xi\|_{L_4}}{\sqrt{N}} \sqrt{\log\left(c \frac{\|\xi\|_{L_4} d}{\sqrt{N}\rho}\right)} \quad (6.9)$$

as long as  $\|\xi\|_{L_4} d / \sqrt{N} \rho \geq c'$ .

Using the constraint that  $\rho/r \geq c\sqrt{s}$ , it is evident from (6.8) that  $s \leq c(L)N/\log(ed/N)$ , and that

$$\frac{1}{s} \geq c_1(L) \frac{r^2}{\rho^2} \gtrsim \frac{1}{\rho} \cdot \frac{\|\xi\|_{L_4}}{\sqrt{N}} \sqrt{\log\left(e \frac{\|\xi\|_{L_4} d}{\sqrt{N}\rho}\right)}.$$

Therefore, to have a ‘legal’ choice of  $\rho$  and  $r$ , we must have

$$N \geq c_2(L)s \log\left(\frac{ed}{s}\right),$$

and

$$\rho \geq c_3(L) \frac{s}{\sqrt{N}} \|\xi\|_{L_4} \cdot \sqrt{\log\left(\frac{ed}{s}\right)}.$$

This naturally leads to the following choices: if the degree of sparsity  $s = d/2^{\ell-1}$  and  $s \geq c(L)N/\log(ed/N)$ , set  $\rho_\ell = r_\ell = \infty$ . The statistical intuition behind this choice is that one is not given enough data to obtain nontrivial information.

If the reverse inequality holds, set

$$\rho_\ell = c(L) \frac{d}{2^\ell \sqrt{N}} \|\xi\|_{L_4} \sqrt{\log(e2^\ell)} \sim_L \frac{d\sqrt{\ell}}{2^\ell \sqrt{N}} \|\xi\|_{L_4},$$

and for that choice  $\rho_\ell$ , the required value of  $r_\ell$  is

$$r_\ell \geq c(L) \|\xi\|_{L_4} \sqrt{\frac{s}{N} \log\left(\frac{ed}{s}\right)} \sim_L \|\xi\|_{L_4} \sqrt{\frac{d\ell}{2^\ell N}}.$$

Finally, let  $\hat{r}_\ell \geq r_\ell$  and recall that  $\lambda_\ell \sim_L \hat{r}_\ell^2 / \rho_\ell$ . Set

$$\mathcal{F}_\ell = \{t : \exists v, |\text{supp}(v)| \leq d/2^{\ell-1}, \|t - v\|_1 \leq \rho_\ell\}$$

to be the set of vectors that are ‘well-approximated’ by  $d/2^{\ell-1}$  sparse vectors. Applying Theorem 4.3, these choices complete the proof of Theorem 1.3.  $\blacksquare$

## 6.2 The tournament SLOPE

The argument used here is similar to the one used for the tournament LASSO, and so we skip most of the details.

In tournament SLOPE one selects  $\beta_i \leq c_0 \sqrt{\log(ed/i)}$  and therefore the corresponding indexing set is contained in

$$V_{\rho,r} = \rho \mathcal{B} \cap r B_2^d = \left\{ v \in \mathbb{R}^d : \|v\|_2 \leq r \text{ and } \sum_{i=1}^d v_i^* \sqrt{\log(ed/i)} \leq \rho/c_0 \right\},$$

where here, as always,  $(v_i^*)_{i=1}^d$  denotes the non-increasing rearrangement of  $(|v_i|)_{i=1}^d$ .

Because  $V_{\rho,r}$  has enough symmetries, one may apply Theorem 6.2, leading to an upper bound on  $r_E$  when

$$\mathbb{E} \sup_{v \in V_{\rho,r}} \sum_{i=1}^d g_i v_i \leq \kappa \sqrt{N} r. \quad (6.10)$$

Also, to estimate  $r_{\mathbb{M}}$  it suffices to ensure that

$$\|\xi\|_{L_4} \mathbb{E} \sup_{v \in V_{\rho,r}} \sum_{i=1}^d g_i v_i \leq \kappa \sqrt{N} r^2. \quad (6.11)$$

Next, one may verify (see Lemma 4.3 in [5]) that if we set  $B_s = \sum_{i \leq s} \beta_i / \sqrt{i}$  and if  $B_s \lesssim r/\rho$ , then  $\Delta(\rho, r) \geq (4/5)\rho$  for centres that are ‘well approximated’ by  $s$ -sparse vectors. Also, for our choice of  $\beta_i$ ,  $B_s \lesssim C \sqrt{s \log(ed/s)}$ . Hence, for a fixed degree of sparsity  $1 \leq s \leq d$ , one has the constraint that

$$\frac{r}{\rho} \geq C_1 \sqrt{s \log(ed/s)} \quad (6.12)$$

for a constant  $C_1$  that depends only on  $c_0$ .

Following the same path used for the tournament LASSO, there is a nontrivial choice of  $\rho$  and  $r$  only when  $s \lesssim_L N / \log(ed/N)$ ; otherwise,  $\rho = r = \infty$  as one would expect. When  $s \lesssim_L N / \log(ed/N)$ , we follow the computation in [5]: let  $s = d/2^{\ell-1}$  and define

$$\rho_\ell \sim_L \|\xi\|_{L_4} \frac{s}{\sqrt{N}} \log \left( \frac{ed}{s} \right) \sim_L \|\xi\|_{L_4} \frac{d\ell}{2^\ell \sqrt{N}}$$

and

$$r_\ell \sim_L \|\xi\|_{L_4} \sqrt{\frac{s}{N} \log \left( \frac{ed}{s} \right)}.$$

Finally, fix  $\hat{r}_\ell \geq r_\ell$  and set  $\lambda_\ell \sim_L \hat{r}_\ell^2 / \rho$ . Applying Theorem 4.3 for the choice of  $\rho_\ell$ ,  $\hat{r}_\ell$  and  $\lambda_\ell$ , and to the hierarchy

$$\mathcal{F}_\ell = \{t : \exists v, |\text{supp}(v)| \leq d/2^{\ell-1}, \Psi(t - v) \leq \rho_\ell\}$$

which completes the proof of Theorem 1.4. ■

## A Additional proofs

The proofs of Lemma 5.9 and Lemma 5.10 are, in fact, the same as in [5] and [8], respectively. The minor modifications to the original proofs are presented in this appendix solely for the sake of completeness and not in full detail.

### Proof of Lemma 5.9

The proof of Lemma 5.9 follows the same path as that of Theorem 3.2 in [5]. Let us begin by examining

$$(*) = \mathbb{Q}_{f,f^*}(j) + \mathbb{M}_{f,f^*}(j) + \lambda(\Psi(f) - \Psi(f^*))$$

in the set  $\{f \in \mathcal{F} : \Psi(f - f^*) = \rho\}$ . If  $\Psi(f - f^*) = \rho$  one should consider two cases. First, if  $\|f - f^*\|_{L_2} \geq r$  then by the triangle inequality for  $\Psi$ , and since  $\mathbb{Q}_{f,f^*}(j) \geq C\|f - f^*\|_{L_2}^2$  and  $\mathbb{M}_{f,f^*}(j) \geq -(C/4)\|f - f^*\|_{L_2}^2$ , we have

$$\begin{aligned} (*) &\geq C\|f - f^*\|_{L_2}^2 - \frac{C}{4}\|f - f^*\|_{L_2}^2 - \lambda\Psi(f - f^*) \\ &\geq \frac{3C}{4}\|f - f^*\|_{L_2}^2 - \lambda\rho \geq \frac{3C}{4}r^2 - \lambda\rho > 0, \end{aligned} \quad (\text{A.1})$$

provided that

$$\lambda \leq \frac{3C}{4} \cdot \frac{r^2}{\rho}. \quad (\text{A.2})$$

If, on the other hand,  $\|f - f^*\|_{L_2} \leq r$ , then  $\mathbb{Q}_{f,f^*}(j) \geq 0$  and  $\mathbb{M}_{f,f^*}(j) \geq -(C/4)r^2$ . Therefore,

$$(*) \geq -\frac{C}{4}r^2 + \lambda(\Psi(f) - \Psi(f^*)).$$

Fix  $v \in \mathcal{B}_{f^*}(\rho/20)$  and write  $f^* = u + v$ ; thus  $\Psi(u) \leq \rho/20$ . Set  $z^*$  to be a linear functional that is norming for  $v$  and observe that for any  $f \in E$ ,

$$\begin{aligned} \Psi(f) - \Psi(f^*) &\geq \Psi(f) - \Psi(v) - \Psi(u) \geq z^*(f - v) - \Psi(u) \geq z^*(f - f^*) - 2\Psi(u) \\ &\geq z^*(f - f^*) - \frac{\rho}{10}. \end{aligned} \quad (\text{A.3})$$

Hence, if  $f^* \in \mathcal{F}_\ell$  and  $f \in \mathcal{F} \cap \mathcal{B}_{f^*}(\rho) \cap D_{f^*}(r)$  then optimizing the choices of  $v$  and of  $z^*$ ,  $z^*(f - f^*) \geq \Delta_\ell(\rho, r)$ ; thus

$$\Psi(f) - \Psi(f^*) \geq \Delta_\ell(\rho, r) - \frac{\rho}{10} \geq \frac{7}{10}\rho. \quad (\text{A.4})$$

And, if

$$\lambda \geq \frac{C}{2} \cdot \frac{r^2}{\rho}, \quad (\text{A.5})$$

we have that

$$(*) \geq -\frac{C}{4}r^2 + \lambda \cdot \frac{7}{10}\rho > 0.$$

In other words, if  $\lambda$  is chosen to satisfy both (A.2) and (A.5),  $f \in \mathcal{F}$  and  $\Psi(f - f^*) = \rho$ , it follows that

$$\mathbb{Q}_{f,f^*}(j) + \mathbb{M}_{f,f^*}(j) + \lambda(\Psi(f) - \Psi(f^*)) > 0.$$

Next, if  $\Psi(f - f^*) > \rho$ , there are  $\theta \in (0, 1)$  and  $h \in \mathcal{F}$  that satisfy

$$\Psi(h - f^*) = \rho \quad \text{and} \quad \theta(f - f^*) = h - f^*.$$

If  $\|h - f^*\|_{L_2} \geq r$ , then by the triangle inequality for  $\Psi$  followed by (A.1),

$$\begin{aligned} (*) &\geq \frac{1}{\theta^2} \mathbb{Q}_{h,f^*}(j) + \frac{1}{\theta} (\mathbb{M}_{h,f^*}(j) - \lambda \Psi(h - f^*)) \\ &\geq \frac{1}{\theta} (\mathbb{Q}_{h,f^*}(j) + \mathbb{M}_{h,f^*}(j) - \lambda \Psi(h - f^*)) > 0. \end{aligned}$$

If, on the other hand,  $\|h - f^*\|_{L_2} \leq r$ , then

$$\begin{aligned} (*) &\geq \frac{1}{\theta} \mathbb{M}_{h,f^*}(j) + \lambda(z^*(f - f^*) - 2\Psi(u)) \\ &\geq \frac{1}{\theta} (\mathbb{M}_{h,f^*}(j) + \lambda(z^*(h - f^*) - 2\theta\Psi(u))) \\ &\geq \frac{1}{\theta} (\mathbb{M}_{h,f^*}(j) + \lambda(z^*(h - f^*) - 2\Psi(u))) > 0, \end{aligned}$$

because  $0 \leq \theta < 1$  and using (A.3).

Now, all that remains is to control  $f \in \mathcal{F} \cap \mathcal{B}_{f^*}(\rho)$  and show that if  $\|f - f^*\|_{L_2} \geq r$ , then

$$\mathbb{Q}_{f,f^*}(j) + \mathbb{M}_{f,f^*}(j) + \lambda(\Psi(f) - \Psi(f^*)) > 0.$$

This follows from (A.1). ■

### Proof of Lemma 5.10

The first part of Lemma 5.10 is identical to Lemma 5.1 from [8], with the trivial modification that the constant  $-C/4$  replaces  $-3C/4$  used in [8]. The second part of Lemma 5.10 was not needed in [8], but its proof follows the same path as Lemma 5.1 from [8].

Set  $r = \hat{r}_\ell$  and fix  $f \in \mathcal{F}$  that satisfies  $\|f - f^*\|_{L_2} \leq r$ . Recall that  $m = N/n$  and that  $\sqrt{n/N} \leq \sqrt{\theta}r/\sigma_4$  for a well-chosen constant  $\theta$  that depends only on  $L$  and  $\tau$ . Set  $U = (f - f^*)(X) \cdot (f^*(X) - Y)$  and observe that

$$\mathbb{M}_{f,f^*} = \frac{1}{m} \sum_{i=1}^m U_i.$$

It follows from the convexity of  $\mathcal{F}$  that  $\mathbb{E}U = \mathbb{E}(f - f^*)(X) \cdot (f^*(X) - Y) \geq 0$ ; therefore,  $\mathbb{M}_{f,f^*} \geq \mathbb{M}_{f,f^*} - \mathbb{E}\mathbb{M}_{f,f^*}$ . Also,

$$Pr(|\mathbb{M}_{f,f^*} - \mathbb{E}\mathbb{M}_{f,f^*}| > t) \leq t^{-1} \mathbb{E}|\mathbb{M}_{f,f^*} - \mathbb{E}\mathbb{M}_{f,f^*}|,$$

and by a straightforward symmetrization argument,

$$\mathbb{E}|\mathbb{M}_{f,f^*} - \mathbb{E}\mathbb{M}_{f,f^*}| \leq 2\mathbb{E}\left|\frac{1}{m} \sum_{i=1}^m \varepsilon_i U_i\right| \leq \frac{2}{\sqrt{m}} (\mathbb{E}|U|^2)^{1/2}.$$

Applying Assumption 2.1, it is evident that

$$(\mathbb{E}|U|^2)^{1/2} \leq \|f^*(X) - Y\|_{L_4} \cdot \|f - f^*\|_{L_4} \leq Lr\sigma_4,$$

and thus

$$Pr(|\mathbb{M}_{f,f^*} - \mathbb{E}\mathbb{M}_{f,f^*}| > t) \leq \frac{2Lr\sigma_4}{t\sqrt{m}} = \frac{2L\sigma_4 r \sqrt{n}}{t\sqrt{N}} \leq \frac{\tau}{3},$$

where we use the fact that  $\sqrt{n/N} \leq \sqrt{\theta}r/\sigma_4$  and select  $t = Cr^2/8$  and  $\theta = \theta(\tau, L)$ . Therefore,

$$Pr(\mathbb{M}_{f,f^*} \leq -(C/8)r^2) \leq \frac{\tau}{3},$$

and with probability at least  $1 - 2\exp(-c\tau^2n)$ ,

$$|\{j : \mathbb{M}_{f,f^*}(j) \geq -(C/8)r^2\}| \geq (1 - \tau/2)n. \quad (\text{A.6})$$

The rest of the argument is identical to the proof of Lemma 5.1 from [8]: let  $\mathcal{H}$  be a maximal separated subset of  $\mathcal{F} \cap \mathcal{B}_{f^*}(\rho) \cap D_{f^*}(r)$  with respect to the  $L_2$  norm, of cardinality  $\exp(c\tau^2n/2)$ , and with the following property: for any  $f \in \mathcal{F} \cap \mathcal{B}_{f^*}(\rho) \cap D_{f^*}(r)$  there is  $h \in \mathcal{H}$  for which

$$\|f - h\|_{L_2} \leq \varepsilon \quad \text{and} \quad \mathbb{E}(f^*(X) - Y)(f(X) - h(X)) \geq 0; \quad (\text{A.7})$$

here  $\varepsilon$  denotes the mesh of the net. The existence of such a separated set is established in [8] (see Lemma 5.3), and one may show that the mesh  $\varepsilon$  is a small proportion of  $r$ .

By (A.6), we have that with probability at least  $1 - 2\exp(-c\tau^2n/2)$ , for every  $h \in \mathcal{H}$

$$|\{j : \mathbb{M}_{h,f^*}(j) \geq -(C/8)r^2\}| \geq (1 - \tau/2)n. \quad (\text{A.8})$$

For every  $f \in \mathcal{F} \cap \mathcal{B}_{f^*}(\rho) \cap D_{f^*}(r)$  let  $\pi f \in \mathcal{H}$  be as in (A.7), and at the heart of the proof of Lemma 5.4 in [8] is that with probability at least  $1 - 2\exp(-c_1\tau^2n)$ ,

$$\sup_{f \in \mathcal{F} \cap \mathcal{B}_{f^*}(\rho) \cap D_{f^*}(r)} |\{j : \mathbb{M}_{f,f^*}(j) - \mathbb{M}_{\pi f,f^*}(j) \leq -(C/8)r^2\}| \leq \frac{\tau n}{2}. \quad (\text{A.9})$$

Combining (A.8) and (A.9), there is an event of probability at least  $1 - 2\exp(-c_2\tau^2n)$  on which for any  $f \in \mathcal{F} \cap \mathcal{B}_{f^*}(\rho) \cap D_{f^*}(r)$  there is a set of coordinate blocks  $(I_j)_{j \in J}$ , of cardinality  $|J| \geq (1 - \tau)n$  and for  $j \in J$ ,

$$\mathbb{M}_{f,f^*}(j) \geq \mathbb{M}_{\pi f,f^*}(j) + (\mathbb{M}_{f,f^*}(j) - \mathbb{M}_{\pi f,f^*}(j)) \geq -\frac{C}{4}r^2.$$

■

## References

- [1] M. Bogdan, E. van den Berg, C. Sabatti, W. Su, and E. Candès. SLOPE-adaptive variable selection via convex optimization. *Annals of Applied Statistics*, 9:1103–1140, 2015.
- [2] R. M. Dudley. *Uniform central limit theorems*, volume 142 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, New York, second edition, 2014.
- [3] V. Koltchinskii. *Oracle inequalities in empirical risk minimization and sparse recovery problems*, volume 2033 of *Lecture Notes in Mathematics*. Springer, Heidelberg, 2011. Lectures from the 38th Probability Summer School held in Saint-Flour, 2008, École d'Été de Probabilités de Saint-Flour. [Saint-Flour Probability Summer School].

- [4] G. Lecué and S. Mendelson. Learning subgaussian classes: Upper and minimax bounds. In S. Boucheron and N. Vayatis, editors, *Topics in Learning Theory*. Societe Mathematique de France, 2016.
- [5] G. Lecué and S. Mendelson. Regularization and the small-ball method I: sparse recovery. *manuscript*, 2016.
- [6] G. Lecué and S. Mendelson. Sparse recovery under weak moment assumptions. *Journal of the European Mathematical Society*, to appear.
- [7] M. Ledoux and M. Talagrand. *Probability in Banach Space*. Springer-Verlag, New York, 1991.
- [8] G. Lugosi and S. Mendelson. Risk minimization by median-of-means tournaments. *preprint*, 2016.
- [9] S. Mendelson. Learning without concentration. *Journal of the ACM*, 62:21, 2015.
- [10] S. Mendelson. Local vs. global parameters—breaking the Gaussian complexity barrier. *manuscript*, 2015.
- [11] S. Mendelson. On aggregation for heavy-tailed classes. *Probability Theory and Related Fields*, to appear.
- [12] S. Mendelson. On multiplier processes under weak moment assumptions. *Geometric Aspects of Functional Analysis - GAFA Seminar notes*, to appear.
- [13] S. Mendelson. Upper bounds on product and multiplier empirical processes. *Stochastic Processes and their Applications*, to appear.
- [14] W. Su and E. Candès. SLOPE is adaptive to unknown sparsity and asymptotically minimax. *The Annals of Statistics*, 44:1038–1068, 2016.
- [15] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B.*, 58:267–288, 1996.
- [16] A.W. van der Vaart and J.A. Wellner. *Weak convergence and empirical processes*. Springer-Verlag, New York, 1996.